the plant journal

SEB Society for Experimental Biology

The Plant Journal (2016) **88,** 648–661 doi: 10.1111/tpj.13274

RESOURCE

Inter-varietal structural variation in grapevine genomes

Maria Francesca Cardone[1], Pietro D'Addabbo[2], Can Alkan[3], Carlo Bergamini[1], Claudia Rita Catacchio[2], Fabio Anaclerio[2], Giorgia Chiatante[1,2], Annamaria Marra[2], Giuliana Giannuzzi[2,4], Rocco Perniola[1], Mario Ventura[2,*] and Donato Antonacci[1,*]

[1]*Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA)-Unità di ricerca per l'uva da tavola e la vitivinicoltura in ambiente mediterraneo, Research Unit for viticulture and enology in Southern Italy, Turi (BA), Italy,*

[2]*Dipartimento di Biologia, Università degli Studi di Bari 'Aldo Moro', Bari, Italy,*

[3]*Department of Computer Engineering, Bilkent University, Ankara TR-06800, Turkey, and*

[4]*Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland*

Received 16 March 2015; revised 12 July 2016; accepted 13 July 2016; published online 16 September 2016.
*For correspondence (e-mails mario.ventura@uniba.it; donato.antonacci@crea.gov.it).

SUMMARY

Grapevine (*Vitis vinifera* L.) is one of the world's most important crop plants, which is of large economic value for fruit and wine production. There is much interest in identifying genomic variations and their functional effects on inter-varietal, phenotypic differences. Using an approach developed for the analysis of human and mammalian genomes, which combines high-throughput sequencing, array comparative genomic hybridization, fluorescent *in situ* hybridization and quantitative PCR, we created an inter-varietal atlas of structural variations and single nucleotide variants (SNVs) for the grapevine genome analyzing four economically and genetically relevant table grapevine varieties. We found 4.8 million SNVs and detected 8% of the grapevine genome to be affected by genomic variations. We identified more than 700 copy number variation (CNV) regions and more than 2000 genes subjected to CNV as potential candidates for phenotypic differences between varieties.

Keywords: *Vitis vinifera* L., table grape, high-throughput sequencing, genomic variation, copy number variation, single nucleotide polymorphism, SRP009057, candidate genes.

INTRODUCTION

Grapevine berries and their derivatives have a large and growing worldwide market as wine, table grapes, raisins, and as products used for human health and cosmetics. Grapevine is also an important system for fruit species, as it can be transformed and micro-propagated via somatic embryogenesis (Velasco *et al.*, 2007) and has sequenced genomes. Two genotypes of Pinot noir (PN; wine grape) have been sequenced and assembled as reference genomes (Jaillon *et al.*, 2007; Velasco *et al.*, 2007). The genome of the Thompson seedless cultivar (TS; table grape) was sequenced and assembled more recently (Di Genova *et al.*, 2014), which enables exploring the genomic differences between wine and table grapes.

Most modern grape varieties resulted from human selection and vegetative propagation with a focus on specific traits of pathogen resistance or crop production. Existing cultivars display a great level of inter-specific variation, which can be investigated to find genes selected by breeding. Interest in the DNA sequence variation in grapevine includes single nucleotide variants (SNVs; e.g. single nucleotide polymorphisms or SNPs), small indels, and structural variations, which include large copy number variation (CNV). SNVs can be rapidly identified by re-sequencing and, in general, their effects on gene functions are relatively easy to detect. Structural variations such as large duplications and deletions need more effort to be correctly characterized. A recent study on structural variation in the PN variety identified duplications in genes responsible for adaptation and response to environmental changes, which are relevant for cultivation needs such as improved resistance against pathogens and tolerance of climate variability (Giannuzzi *et al.*, 2011).

Despie the importance of CNVs, our understanding of the most prevalent contributors to CNVs in plants is still far from being well explored. Genome data obtained from different plant species revealed the plasticity of plant genomes, and

these findings led researchers to extend the concept of the 'pan-genome', first described for bacteria (Tettelin *et al.*, 2005), to plant species. The pan-genome has been defined as the ensemble of a core portion, present in all the individuals, and a dispensable portion, not present in all individuals. The latter has been considered not essential for survival. Nevertheless, the high level of structural variations found in many plant genomes suggests that dispensable genomes may have an important role in shaping genome structure (Marroni *et al.*, 2014). High levels of CNVs have been found distributed throughout many plant genomes contributing to phenotypic variation associated with phenotypic traits (Hurwitz *et al.*, 2010; Cao *et al.*, 2011a,b; Haun *et al.*, 2011; Saintenac *et al.*, 2011; Yu *et al.*, 2011; Zheng *et al.*, 2011; Chia *et al.*, 2012; McHale *et al.*, 2012). This has led to an increasing interest in studying all forms of genomic variation in plant genomes (Żmieńko *et al.*, 2014).

We combined high-throughput sequencing (HTS) with array comparative genomic hybridization (CGH), fluorescent *in situ* hybridization (FISH) and quantitative PCR (qPCR) to create a comprehensive map of genomic variations in four table grape genomes. We sequenced and compared the four table grape cultivars (cv): Autumn royal (AR); Italia (It); Red globe (RG); and TS; with the PN genome as the reference (inbred line PN40024). We found that 8% of the grapevine genome is affected by genomic variations and is characterized by a high level of plasticity detected as inter-varietal-specific CNVs and SNVs – the latter corresponding to an average of 1 SNV every 100 bp (4.8 million detected SNVs). We performed an in-depth analysis of gene content of polymorphic regions and detected varietal CNVs in genes involved in aromatic compound biosynthesis and metabolism related to aromatic berry flavor. Likewise, notable genomic variation differences were found for genes playing roles in stress response to both biotic and abiotic stresses, such as an S-locus lectin protein kinase and an NADH dehydrogenase subunit 7, completely deleted and highly duplicated, respectively, in the four table grape genomes we analyzed. Overall, we created comprehensive variation maps that allow for the identification of genes and/or gene families as putative functional candidates for important traits in grapevine cultivation. Polymorphic genes described in this paper require further analysis on a larger sample size to validate their role in such trait determination; however, these data may represent a landmark that can be used to develop genetic tools for breed selection programs.

RESULTS

Sequencing and variant calling

We sequenced the four table grape cultivars: AR; It; RG; and TS; using 76 nt paired-end reads. Sequence coverage ranged from 13× to 19×. Sequence reads were aligned against the PN40024 reference genome.

To generate a more accurate SNP call set, we pooled the data of each single variety according to the best practices guidelines (McKenna *et al.*, 2010). We identified a total of 4 478 098 SNPs and 262 395 indels after applying quality thresholds as described by McKenna *et al.* (2010; Table S1). We then compared SNP and indel call sets among the four varieties to characterize shared and cultivar-specific sequence variants (Table 1). We also compared the SNPs found in these four varieties with those reported in the literature by matching our list of SNPs with that published by Di Genova *et al.* (2014); 746 560 SNPs were found in common in both lists, which represented 16.7% of the SNPs found in the present work pooling four varieties and 57.7% of the SNPs found in TS by Di Genova *et al.* (2014). As the previous work was done only with TS, while our work analyzed a pool of four varieties, as in our case, we focused a second comparison on TS. Among our SNP calls, 3 117 684 belong to the TS cv, and 731 000 of them (~23.4%) were present also among the 1 292 709 SNP calls presented by Di Genova *et al.* (~56.5%). These data confirmed the higher sensitivity of our pooled-variety sequencing method in detecting a comprehensive list of SNPs.

Next, we used the SnpEff tool (Cingolani *et al.*, 2012) to predict the effects of the alternative allele for those variants mapping within coding regions. We detected 5136 variants that add or remove stop codons, therefore potentially altering the length of the coded protein (Table S2). Among those variants, 539 were previously reported (Di Genova *et al.*, 2014). We further Sanger-sequenced randomly selected SNPs for genotype validation. In PN40024, 100% of the predicted SNPs were validated, whereas the percentage of validation for the four table grape varieties ranged between 78% and 84% (Table S3). The majority of the unvalidated SNPs were predicted as heterozygous, but were found homozygous concordant with the reference genome. They were then counted as false-positives, even if these results may still be biased, as the step of PCR amplification preliminary to the validation method could cause allelic drop out.

Based on the read-depth analysis [whole-genome shotgun detection (WSSD); see Methods], we generated a duplication and deletion map for each variety (Figure S1). We calculated absolute copy number (CN) values and identified deletions together with duplications that were essential to perform a multi-varietal comparison for CNV identification (Alkan *et al.*, 2009).

Whole-genome shotgun detection analysis revealed similar percentages of duplication (average 16%) and deletion (average 3%) in the four table grape varieties and in the PN40024 reference. We further compared duplicated/deleted regions among the five varieties, and found 26.13% of the grape genome duplicated in at least one variety (Table S4). In particular, 18.58% of the grape genome was predicted to be composed of segmental

Table 1 Summary results of genomic variations

Shared polymorphisms (# in millions)

	Total SNPs	/	Autumn royal	Italia	Red globe	Thompson seedless	Total indels
Autumn royal	2.44	/	/	0.13	0.12	0.14	0.16
Italia	2.88	/	1.72	/	0.13	0.14	0.19
Red globe	2.61	/	1.56	1.81	/	0.14	0.17
Thompson seedless	3.12	/	1.91	1.96	1.87	/	0.20

Varieties-pairs shared WSSD (Mbp)

	Total duplications	PN40024	Autumn royal	Italia	Red globe	Thompson seedless	Total deletions
PN40024	79.48 (16.35%)	/	3.04	2.88	2.50	3.33	14.37 (2.96%)
Autumn royal	71.39 (14.68%)	49.26	/	8.02	6.30	8.94	17.45 (3.59%)
Italia	77.78 (16.00%)	52.21	57.28	/	6.40	7.89	14.26 (2.93%)
Red globe	75.31 (15.49%)	48.68	53.78	58.64	/	7.53	13.99 (2.88%)
Thompson seedless	75.87 (15.60%)	50.50	57.20	59.22	56.35	/	18.03 (3.71%)

Varieties-pairs shared CNV (Mbp)

	Total OVER	/	Autumn royal	Italia	Red globe	Thompson seedless	Total DOWN
Autumn royal	7.49 (1.54%)	/	/	5.23	4.11	5.64	12.09 (2.49%)
Italia	7.70 (1.58%)	/	3.72	/	3.74	4.25	8.66 (1.78%)
Red globe	9.25 (1.90%)	/	4.41	4.09	/	4.33	8.70 (1.79%)
Thompson seedless	8.62 (1.77%)	/	3.71	3.48	3.78	/	11.10 (2.28%)

Shared SNP: millions of SNPs (below the diagonal) and indels (above the diagonal).
Varieties-pairs shared WSSD: duplication (below the diagonal) and deletion (above the diagonal) coverage with respect to the average CN status. In brackets, percent (%) of the genome duplicated or deleted.
Varieties-pairs shared CNV: over (below the diagonal) and down (above the diagonal) coverage with respect to the Pinot noir CN status (L2R positive or negative).
CNV, copy number variation; SNP, single nucleotide polymorphism; WSSD, whole-genome shotgun detection.

duplications (SDs) that are common to at least two varieties (shared SDs), while 7.54% were unique duplications as they were found in only one variety. Notably, 1017 regions (1.72% of the genome) were found duplicated in all four table grape varieties. Likewise, we found about 9% of the grape genome deleted in at least one of the five analyzed varieties: 5.14% were found as unique deletions, while 3.86% contained deletions common to at least two varieties. Only 0.55% (210 regions) represented common deletions to the four table grapes (Tables 1 and S4).

Next, to detect CNVs differentiating the varieties, we applied an *in silico* digital CGH approach on the whole genome similar to an algorithm described to characterize CNVs within human genomes (Sudmant *et al.*, 2010). We calculated the log2 ratio (L2R) between the CN of a region in one of the four sequenced cultivars and the CN of the matching region in the reference genome (Sudmant *et al.*, 2010). As this approach has been implemented only on humans and mammals, but not on a plant genome, we tested different L2R thresholds and different cut-off windows (number of consecutive windows matching the L2R threshold). We then chose the parameters that allowed for the identification of larger copy number variant regions (CNVRs; > 10 kbp), while reducing the putative false-positive calls. We used a threshold of L2R 0.25, which means differences in CN of at least 20%, thus allowing for easier identification of polymorphisms in regions where differences of 20% could be significant (e.g. CN = 20 versus CN = 16; Appendix S1-§*1.1 Digital CGH*).

Compared with the PN40024 reference, we found 1.5–1.9% and 1.7–2.5% of the genome in each variety as amplified or deleted, respectively (Table 1). We detected a total of 746 CNVRs (>10 kbp) overlapping across the four varieties: 310 in It, 318 in RG, 355 in TS and 350 in AR (Figure 1 and Table S5), which are equally distributed between gains and losses of paralogous copies. This corresponds to a percentage of variant genome, ranging from 3.35% in It to 4.05% in TS.

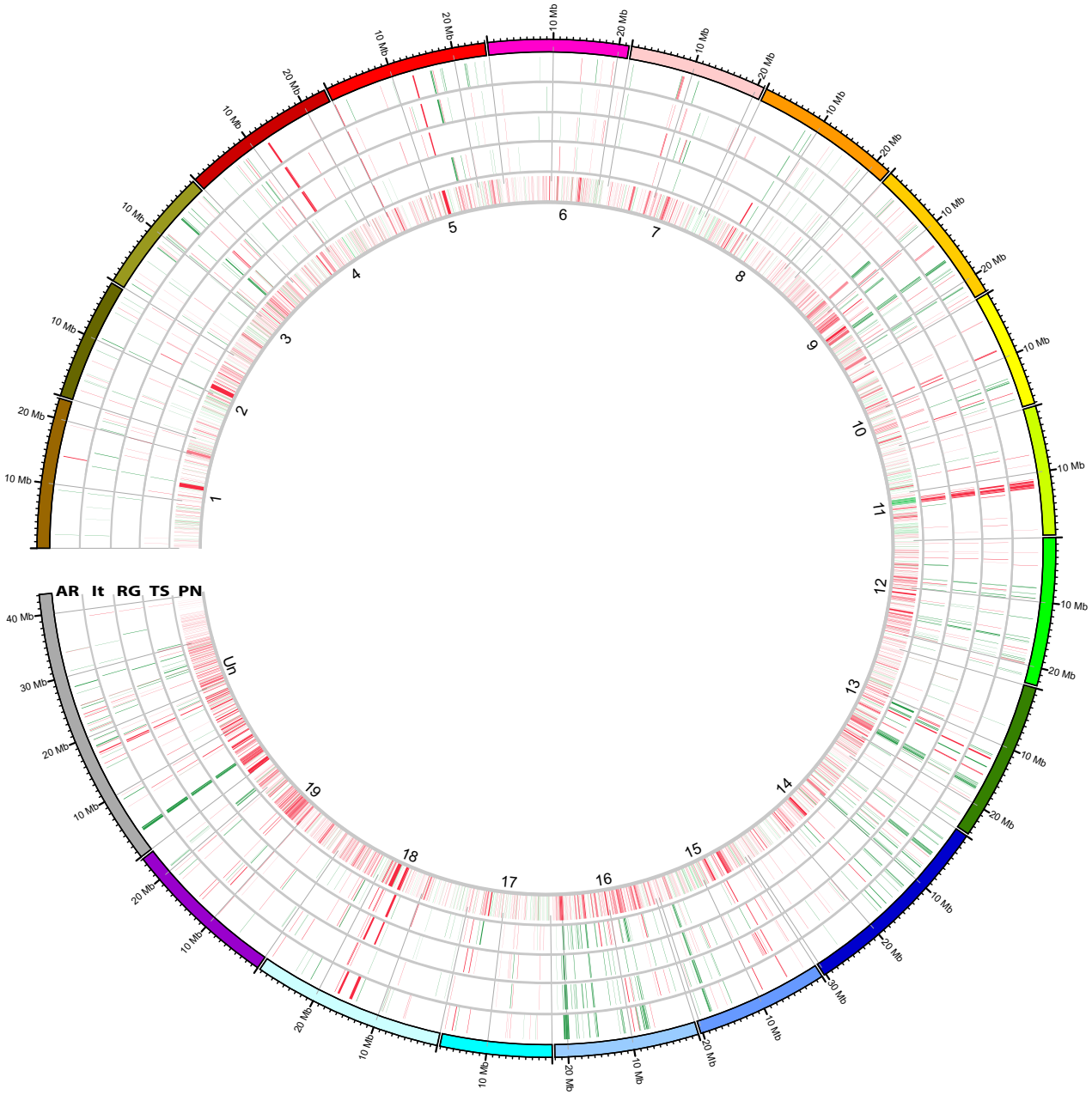

Figure 1. Circular representation of dCGH (comparative genomic hybridization) data in four varieties of grapes.
Deleted (green) and duplicated (red) regions, detected by dCGH with respect to the PN40024 genome, were graphically highlighted in four circular representations of the genomes of the analyzed grapes varieties. The external colored circle represents the 19 *Vitis vinifera* L. chromosomes and the 'unknown' chromosome (that collects regions sequenced but unassigned to a specific chromosome). Chromosome name is reported and vertical gray lines delimit the start and end of each chromosome. The internal circles tag copy number variations (CNVs) found in each region, reported in order, in Autumn royal (AR), Italia (It), Red globe (RG) and Thompson seedless (TS). The inner circle represents the whole-genome shotgun detection (WSSD) map in the PN40024 reference genome, where the red tags indicate segmental duplications (SDs) and green tags indicate deletions. As an example, the schema shows a big polymorphic region (CNV) in chr11, whose duplication with respect to the PN40024 is shared between all the varieties. Likewise, a shared deletion can instead be observed at the end of the chr16.

In each of the four varieties, about 35% of these regions were large CNVs greater than 50 kbp, and 10% were greater than 100 kbp. Out of the 746 CNVRs, 335 CNVRs were uniquely identified in one variety, while 64 were found in all four table grape genomes with respect to the reference (Table S5). Overall, our inter-varietal comparison revealed that about 8% of the grapevine genome is characterized by CNVs.

We further compared CNV calls with WSSD results and checked if CNVRs overlapped with regions found

duplicated or deleted in the reference genome (WSSD-positive regions). Notably, 46% of the CNVRs were also positive to the WSSD analysis on the reference genome (both duplicated or deleted), and about 41% of CNVs mapped in regions duplicated in the reference genome, while the other 5% matched with regions deleted in the PN40024 genome (Figures 1 and S2).

Methods for genome-wide detection of CNVR have only been implemented, until now, with humans and mammals (Sudmant *et al.*, 2010; Bickhart *et al.*, 2012). Only a few studies reported similar approaches on plant genomes (Żmieńko *et al.*, 2014), and there are to the best of our knowledge no previous genome-wide studies on CNVRs in grapevine. For this reason, we performed experimental validation to confirm individual CNVs using array CGH, FISH and qPCR. Array CGH revealed aberrations in about 2% of the grape genome for each of the analyzed cultivars. We found 102 aberrant regions in It, 121 in RG, 124 in TS, and 166 in AR (Figure S3; Table S6). Compared with the digital CGH approach, array CGH detected more deletions than amplifications, as expected due to the signal saturation biases within duplicated regions.

Data from array CGH were compared with those from digital CGH, and percentages of concordance between array versus digital calls were calculated (Appendix S1-§*1.2 Digital CGH versus Array CGH*). We could validate about 25–30% of digital calls using array experiments. This level of validation was probably due to both the low level of resolution of array CGH with respect to the *in silico* approach and to the technical limit of array assay. For example, array platforms tend to suffer reduced sensitivity in the detection of amplification (Alkan *et al.*, 2011). Indeed, the validation rate also reached 37.5% if only deletions were considered. In addition, we account that the error rate is related to the quality level of the draft of the reference genome. In human, that could be considered the most complete genome, Sudmant *et al.* (2010) found that the false detection rate at 1 kbp resolution (size of the windows) with a coverage >8 × is approximately 1–3%. Both deepening the coverage and focusing on larger CNVs could lower the rate. According to this, we reached a higher coverage for each of the sequenced genomes and searched for CNVRs > 10 kbp.

To further confirm variety-specific copy gains and losses, we performed FISH on interphase nuclei (Figure 2) using 43 BAC clones as probes (Table S7). We successfully tested 34, 37, 27 and 33 regions on AR, It, RG and TS nuclei, respectively. The validation rate was variable among the varieties, ranging from 58% in TS, 63% in RG, to 68% in AR and It; 16 BACs showed reliable results in all four varieties and the validation rate of this subset supports the percentages of the whole pool. While, to the best of our knowledge, no FISH data are available on plant genomes, a similar validation rate, using FISH assays, has

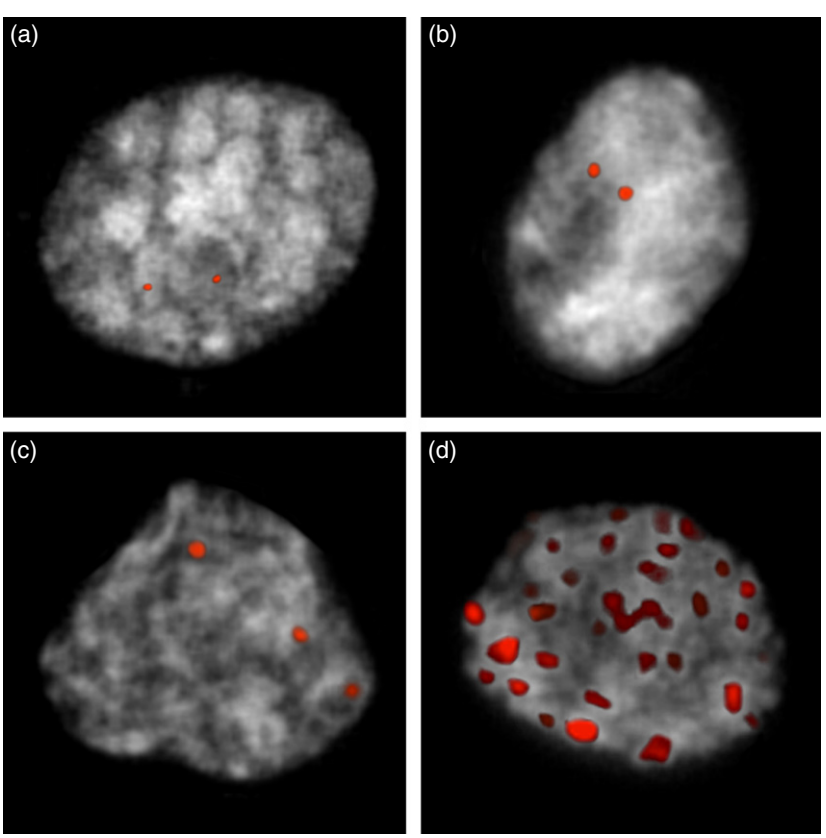

Figure 2. The four main hybridization patterns observed on *Vitis vinifera* L. interphase nuclei are shown.
They exemplify the fluorescent signals produced with probes containing: (a) a single region, (b) a tandem duplication, (c) an interchromosomal duplication and (d) a hyperexpanded (tandem and interspersed) duplication.

been described for mammals (Liu *et al.*, 2010; Hou *et al.*, 2011; Bickhart *et al.*, 2012). Seven BAC clones showed a complex hybridization pattern revealing highly duplicative and polymorphic levels (Figure S4), confirming the *in silico* prediction that showed multiple subregions with different CN status (Table S8).

Further qPCR assays were designed to confirm 21 genic regions predicted to be polymorphic among the four sequenced varieties. In particular, we selected three regions predicted as constantly diploid and 18 predicted to be highly polymorphic among the varieties (Table S9). CN was estimated using the relative standard curve method, comparing with an endogenous reference gene arbitrarily taken as constantly diploid, the fructose-6-phosphate-2-kinase. Among the 21 tested genes, 17 were validated by qPCR assays, while two gave unreliable amplifications due to the presence of non-specific PCR products and the other two genes were found not confirmed. For the 17 validated genes, we found a robust correlation between the *in silico* CN predictions, and those from the qPCR linear regression analysis indeed revealed a correlation coefficient R of 0.802 ($P < 0.0001$) if all the varieties were considered together, and ranging between 0.742 for RG and 0.869 for AR, if the sequenced varieties were considered one by one.

As a further corroboration of the reliability of the absolute CN *in silico* predictions, we calculated the linear regression among all the $CN_{in\ silico}$ with respect to the CN_{qPCR}. The function describing the regression of the data was found as follows: $CN_{in\ silico} = 1.001 * CN_{qPCR} - 0.540$ (Figure S5; Appendix S1-§*2 Validation of CN estimation by qPCR*).

Functional analysis of CNVs genes

To find polymorphic genes, we compared the CNVRs with *Vitis vinifera* L. gene annotation and searched for specific functional category among polymorphic regions. Based on the gene content, the 746 large CNVRs were decomposed in 3678 subregions (Table S10). For each subregion, CN was calculated as the average of the CN of the non-overlapping windows of 1 kbp unmasked sequence (KbUS) mapping in that region. In addition, gene name, mapping and functional annotation were reported.

We mapped genes with specific functional category annotation to 2029 out of 3678 subregions (55%). The most abundant category included genes involved in primary metabolism (~27%), especially genes with protein metabolism and amino acid metabolism functions, such as members of tRNA synthases or tRNA-ligases. We also found genes involved in stress response, such as the members of chitinase family. Similarly, almost 16% of the polymorphic regions included genes involved in signaling pathways. Furthermore, we found many of the polymorphic genes that belong to large and well-known gene families and superfamilies – such as those involved in terpenoid and phenyl propanoid metabolism (which account for about 6% of the polymorphic genes), genes belonging to the *MYB* transcription factors superfamily involved in the regulation of different pathways, R-proteins and other disease-resistance proteins of the *NBS-LRR* family (about 10%), genes that belong to the *CYP450* superfamily, and genes belonging to the *EXP* family involved in berry quality determination. Interesting these multigene families are involved in many important aspects related to the berry quality or to the ability to respond to environmental changes, so our results suggest that CNVs in multigene families could explain the great phenotypic variability existing in the Vitis genus. As an example, we found the gene GSVIVT01034920001 that codes for an expansin A4 protein showed a CN = 4 in the reference genome and a CN >10 in all the table grapes. Recently, expression data of these genes revealed they are finely modulated during fruit growth and maturation and, thus, have an important role in processes critical to determining berry quality (Dal Santo *et al.*, 2013). The economic importance of grapevine is greatly influenced by the quality of its berry. This is especially true for the table grape as berries are the final product; hence, the polymorphisms found in a member of the *EXPA4* family support the important role of these genes in berry development.

We also found many transposable element genes that map to both amplified and deleted regions. Indeed, this category was the second most abundant category, reaching almost 18% of polymorphic regions. We focused on gene content in the most polymorphic amplified regions named 'hyper-duplicated regions' (CN > 10). In addition to many genes that code for transposable elements, we mapped genes belonging to some of the most important gene families in grapevine, such as the *TPS* family, *CYP450* family and the *CC-NBS-LRR* defense gene families involved in determining important quality aspects of grapes (e.g. flavor content), or in stress and environmental responses (Table S8). In particular on chr18_random (3 834 790: 3 911 181), we mapped six different genes (GSVIVT01036325001; GSVIVT01036327001; GSVIVT01036328001; GSVIVT01036331001; GSVIVT01036332001; GSVIVT01036333001) that are all annotated as copies of germacrene D-synthase, a member of the *TPS* family (Martin *et al.*, 2010). This region was shown to be duplicated in all analyzed grape genomes, with the It genome showing the highest CN (confirmed by FISH assay), with respect to both the PN40024 reference genome and the other table grape genomes under analysis. Notably, we found an *NADH dehydrogenase subunit 7* (GSVIVT01004966001), a gene involved in the plant vigor, showing CN = 5 in the PN40024 genome, while CN > 30 in the four table grapes. Amplification in this region with respect to the reference genome was confirmed by array CGH assays. Additionally, we observed

that five out of the seven BAC clones revealing highly duplicative and polymorphic levels in FISH assays contained hyper-duplicated regions (Figure S4; Table S8).

Similarly we focused also on regions affected by 'complete' deletion in one or more genomes under study (showing $0.01 < CN < 0.51$ at least in one of the analyzed varieties). We found only 212 such regions among the 3678 polymorphic regions. Among these 212 regions, notably, five regions were found to be completely deleted in all the four table grapes, while PN40024 showed a normal disomic CN in the same regions. Three of these regions contain gene models: GSVIVT01008378001 at chr17: 2,692,174-2,698,467 (an ATP synthase beta chain 2, mitochondrial involved in respiratory chain phosphorylation), VIT_15s0046 g00800 at chr15:17,745,178-17,748,265 (an S-locus lectin protein kinase involved in the signaling pathway), and a transposable element at chr9:20,734,461-20,735,459. Because the two genes are both involved in multiple pathways related to stimulus response, it is difficult to make assumptions on the phenotypic effect of these deletions. Specific functional studies could unveil a possible involvement of these genes in response to environmental changes. The finding of a transposable element in one of these regions may also support a role in mediating gene deletions (Marroni *et al.*, 2014).

We partitioned the polymorphic regions into specific subclasses based on common or unique phenotypes of the four varieties. Next, we searched for significant functional category enrichment and for genes involved in specific pathways related to the phenotypic traits, such as seed content, berry size and aroma compound.

Seedless versus seeded. Among the four sequenced cultivars, the TS and AR cultivars are seedless, while the It and RG cultivars are seeded, like the PN used as the reference. We searched for polymorphic regions with opposite CN status in the seedless varieties with respect to the seeded varieties, taking into account the small sample size and the putative influence of shared ancestry among the seedless varieties. We found 175 regions with CNVs common to the seedless variety and absent in the seeded ones. Next, we searched for specific functional annotation enrichment. We found many regions that contain genes involved in the response to stimulus or hormone signaling. As an example, we found six genes implicated in abscisic acid (ABA) metabolism, which has as important role in drought and other stress responses. All of these genes were found deleted in seedless cultivars. In contrast, the other two genes – GSVIVT01025701001 and GSVIVT01025700001 – coding for *EIN2* (ethylene insensitive 2) involved in ABA and ethylene signaling were mapped in a region amplified in the seedless cultivars. Notably, ABA and ethylene are two key hormones inducing response to abiotic stress in plant, for example one of the mechanisms activated in response to water stress is ABA-mediated and it starts in seeds (Fujita *et al.*, 2013). We could speculate that CNVs found in these genes may influence the response to water stress of seedless varieties, in particular deletion of ABA genes could led to the activation of other mechanisms in response to drought stress like those mediated by *EIN* genes. Another gene of interest was GSVIVT01017620001 coding for the auxin response factor-2, which showed a highly polymorphic status in the analyzed varieties with $CN > 6$ in the seeded and $CN = 3$ in TS and AR. Auxin response genes have been described as regulator, in combination to gibberellins response factors, of fruit set development. However, the molecular mechanisms by which these hormones mediate fruit set initiation are not well established. In particular, fruit set is initiated only after two sequential events, pollination and fertilization, concurrent with changes in the levels of endogenous plant hormones (Jung *et al.*, 2014), but in seedless varieties early in the development embryo abortion occurs leading to seed growth stop. It is plausible that polymorphisms in genes coding to auxin and hormone response could be related to the different response to hormonal signaling in seedless varieties. In addition, we discovered genes coding for transcription factors belonging to important regulatory factor families, such as the *MYB/KANADI* and the zinc finger families, which were deleted in seedless varieties with respect to the seeded ones. The seedless trait is correlated with the berry size, as seedless varieties have usually smaller berry size. Indeed molecular breeding programs aim to obtain seedless variety with big berry size. Among the analyzed varieties in the present study, the TS has a small berry size, while AR reaches larger sizes during berry ripening without hormone treatments. For this reason, we further checked among polymorphisms common to TS and AR for regions with opposite CN status in TS with respect to AR. Noteworthy, we found only two regions with this feature, and only one contained a gene named GSVIVT01018696001 duplicated in TS and deleted in AR that is involved in cell growth and death. Interrogation of grapevine expression atlas (Fasoli *et al.*, 2012) and VITIS Co-expression Database (VTB: http://vtcdb.adelaide.edu.au/Home.aspx) revealed that this gene is specifically expressed in the development stage. In our opinion this gene could represent a good candidate to further investigate aiming to identify genes involved in berry growth and development.

Small versus big berry size (TS versus all others). An important trait in table grapes is berry size, which is related to seedlessness trait. Among the analyzed varieties, TS has the smallest berry size. Therefore, we searched for functional category enrichment in polymorphic regions that are unique in this cultivar. More than 480 polymorphic regions were found to be unique in TS. Among these, many are

involved in cellular and primary metabolism. We also focused our attention on genes involved in *CYP450*-mediated metabolism and auxin/hormone signaling as mechanisms already reported as involved in berry growth (Doligez *et al.*, 2013), and found 10 genes belonging to the *CYP family* and eight genes involved in hormone response. Among these we found the gene GSVIVT01009865001 at chr18:11920505-11928962 that codes auxin response factor-5 (*ARF5*), which was found duplicated in all the analyzed varieties and in the reference, but showed a higher CN in the TS genome and was recently mapped to a quantitative trait locus associated to berry weight and traits (Doligez *et al.*, 2013). We also found CNVs in the gene AUXEFF (GSVIVT01030905001) involved in auxin transport, two genes (GSVIVT01005915001 and GSVIVT01005917001) belonging to the expansin family involved in the auxin signaling pathway, and two genes (GSVIVT01001405001 and GSVIVT01001406001) that code for a GIGANTEA protein implicated in flower development at chr18:1867580 7:18702086. Berry growth and development in grapevine relies on a wide range of control systems, including an intricate network of interactions between all classes of known plant hormones. Our knowledge about the molecular interactions of these different classes of hormones in the ripening process is still imperfect (Böttcher, 2012). Our findings highlight new putative candidates that deserve further functional study to assess this topic, and to understand their role in berry growth and development.

Moreover, we uncovered an enrichment of genes involved in transport overview pathways, such as *PIP2B*, which codes a member of the aquaporin gene family and two genes coding for ABC transporters. Interestingly, Doligez *et al.* (2013) recently discovered new quantitative trait loci for berry weight or seed traits in grapevine, and mapped in these quantitative trait loci genes related to cell wall modifications, water import, auxin and ethylene signaling, transcription control, and organ identity. Likewise, recently, genes involved in the same pathways have been found differentially expressed in grapevine seedless segregants during berry development (Muñoz-Espinoza *et al.*, 2016), thus our results support the hypothetical role of these genes in berry weight and development.

Aromatic versus neutral (It versus all others). Among the analyzed varieties, only the It cultivar is considered to be aromatic. Thus, we investigated the gene content in polymorphic regions of the It genome with respect to all other genomes in order to find candidate genes for this particular trait. We found 346 such CNVRs. Besides enrichment of genes involved in primary metabolism and signaling/stimulus response, we found 15 regions that contain genes for secondary metabolism, especially terpenoid and flavonoid biosynthesis, for example, the germacrene D-synthase described above as mapped in a hyper-duplicated region. The terpenoid pathway is important in the production of fragrance and aroma constituents of flowers and fruits (Martin *et al.*, 2010). Grape genome sequencing has revealed 124 genes related to the terpenoid pathway all organized in gene families originated through duplication events (Velasco *et al.*, 2007). Recently, it has been proved the direct involvement of VvDXS gene in determining the muscat flavor in grapevine (Battilana *et al.*, 2011). Nevertheless, the muscat flavor is not the only flavor determining the aromatic aspect in grapevine, and no significant genetic association was detected to distinguish between aromatic and muscat-flavored fruited varieties, neither to explain flavor intensity variation within the aromatic and muscat groups (Emanuelli *et al.*, 2010). Quantitative or qualitative factors responsible for the neutral to aromatic and aromatic to muscat flavor transition are still unclear. CNVs in TPS genes, such as for the germacrene D-synthase, could be related to the different aromatic component among varieties, similar to variations of terpenoid profiles effected by TPS CNV in other plant species (Hall *et al.*, 2011; Roach *et al.*, 2014).

DISCUSSION

Genome data of different plant species are revealing considerable plasticity, variability and complexity of plant genomes. Despite the important role that variable regions have in the adaptation and evolution of plant genomes, they are still not well characterized. Combining HTS with array CGH, FISH and qPCR assays on four table grape genomes and comparing the data with the reference genome of the PN40024 inbred line, we depicted a detailed inter-varietal atlas of genomic variations in the grapevine genome. Our approach used algorithms specifically designed to predict absolute CNs and characterize SDs.

We identified about 4.8 million high-quality SNVs (SNPs and indels) – about 1 SNV to every 100 bp of grape genome. Previously reported data on SNP detection in grapevine (Lijavetzky *et al.*, 2007; Dong *et al.*, 2010) showed a higher frequency of SNPs in the genome. The larger number of varieties studied in these works may explain the higher SNP detection rate, although in both of these studies only a small portion of the genome was analyzed. Moreover, the differences we observed in SNVs calls in TS when comparing our calls with previous work (Di Genova *et al.*, 2014) were most likely due to the pooling strategy of the sequences of the four genomes and the genomic differences in the two plants sequenced.

Further, we searched for increase/decrease of read-depth coverage comparing the four table grapes with the PN40024 reference genome, and revealed that deletions and highly identical SDs characterize approximately 9 and 26% of grape genome, respectively. Notably, shared deletions and SDs among the four table grapes characterized 0.55 and 1.72% of the genome, respectively, showing the

high plasticity of the grapevine genome among plant genomes (Żmieńko *et al.*, 2014). Interestingly, we compared our SD map on the reference genome with previously reported duplicated regions (Giannuzzi *et al.*, 2011) and showed 50% concordance. It is likely that differences between sequencing methods (next-generation sequencing versus Sanger) and/or call criteria could account for the concordance rate. Comparison among different sequencing-based approaches to detect structural variations already revealed that none of these methods could be considered comprehensive (Alkan *et al.*, 2009, 2011).

Overall, we identified 746 large CNVRs (> 10 kbp) representing about 8% of the genome (Figure 1), with half of them overlapping with SDs (Figures 3 and S2), thus confirming SDs as hotspots for CNV formation (Sharp *et al.*, 2005; Alkan *et al.*, 2009; Marques-Bonet *et al.*, 2009). We also found many transposable elements among deleted polymorphic regions, similar to *Arabidopsis thaliana* (DeBolt, 2010), also supporting the important role of transposon movement in mediating deletions (Morgante *et al.*, 2007; Marroni *et al.*, 2014). Focusing on gene content of plastic regions common to the four table grape genomes, we detected duplication for some well-known multigene families, such as the *MYB* transcription factors, the *TPS* and *EXPA4* (Malacarne *et al.*, 2012) families involved in the grape quality as related to the anthocyanin synthesis, flavonoid or terpenoid metabolism, berry size, maturation

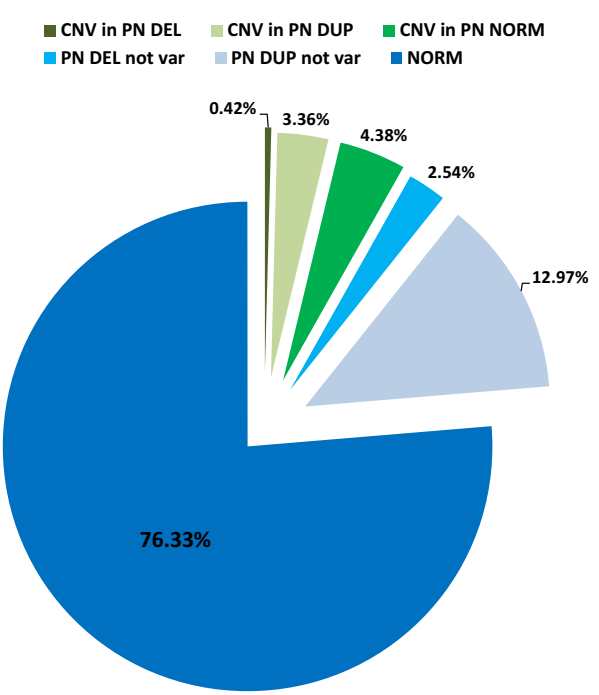

Figure 3. Graphical representation of the results of the comparison between digital comparative genomic hybridization (CGH) and whole-genome shotgun detection (WSSD) analysis. Slice sections describe the percentage of overlap between the copy number variations (CNVs) and the WSSD regions found in the PN40024 reference genome.

and seed formation. Similarly, we found high variability in *NBS-LRR* genes, which are involved in stresses and environmental responses. We hypothesized this as the result of both specific grapevine adaptive responses and human breeding and domestication practices (Matus *et al.*, 2008). Indeed, high levels of duplication ensure the variability of defense genes, and such variation is advantageous in the face of changing environmental conditions. In addition, CNV genes involved in biotic and abiotic resistance could explain the different adaptation to respond to external environmental stresses of one variety with respect to another.

By inter-varietal comparison, we identified candidate genes for specific traits, such as berry flavor, in the It cultivar. The gene for the germacrene D-synthase belonging to the *TPS* superfamily, for example, was found amplified in all the analyzed varieties, but showed a higher CN in It, which is the only variety showing aromatic flavor. Terpenoids contribute distinctively to the sensorial character of aromatic grape varieties; attributes like floral, fruity and citrus can be assigned to a variety depending on the different type and amount of mono- or sequiterpenoid compounds (Martin *et al.*, 2010; May *et al.*, 2013). CNVs in *TPS* could be related to differences in the metabolic pathways of this compound and contribute to differences in the aromatic flavor of one variety with respect to another one. In this context, the polymorphisms found in the germacrene D-synthase and other *TPS* genes represent good candidate genes and deserve further investigation.

Similarly, we found polymorphisms related to seedlessness and berry size in genes involved in auxin/hormone response, berry growth and development. Noteworthy, association mapping studies and expression data revealed candidate genes for berry weight involved in such similar pathways (Muñoz-Bertomeu *et al.*, 2006; Doligez *et al.*, 2013; Muñoz-Espinoza *et al.*, 2016). Our present results support the proposed role of these pathways in such complex traits and highlight new hypothetical candidate genes for further investigation. As future perspective, co-expression and network analysis on those genes could reveal the molecular mechanisms involved in seedlessness and berry traits.

The possibility to calculate the absolute CN of each region allowed for the identification of hyper-duplicated regions. Noteworthy, a member of an NADH dehydrogenase gene family showed CN six times higher in the table grapes than in the reference genome. These genes are involved in cellular homeostasis and oxide-reduction processes, and play a key role in the regulation of plant growth and stress responses (Fujita *et al.*, 2006). Plant productivity, and thus plant vigor, is related to the ability to respond and adapt to environmental stresses. In viticultural terminology, the rate of shoot growth or elongation over time is referred to as vigor, which is influenced by

genotype (species, cultivar and rootstock), and other environmental and cultural aspects. Some cultivars (e.g. TS) or rootstocks (e.g. 1103 Paulsen) are considered to be vigorous, whereas others are thought to be intermediate (e.g. Cabernet Sauvignon, Syrah, Chardonnay, Teleki 5BB or SO4 wine grapes) or weak (e.g. Gamay, 101-14 Mgt or Riparia Gloire rootstocks; Keller and Tarara, 2010). Goff and colleagues recently proposed that efficient energy metabolism and stress response mechanisms are important factors in heterosis, and that plasticity of genes involved in these pathways could explain why heterotic plants are more fit than their corresponding inbred lines (Goff, 2011; Goff and Zhang, 2013). The lower CN for the NADH dehydrogenases that we found in the reference genome could be related to the lower vigor of PN40024 as it is an inbred line.

Noteworthy, we found many unknown or unannotated polymorphic genes, which in fact depend on the low level of gene annotation that still exists for the grapevine genome, even though many efforts are still in progress on this aspect. Further studies on these genes using genotype–phenotype association studies or expression and co-expression data could help to understand their function and improve the functional annotation of the grapevine genome.

Overall, our data suggest that the entire grape genome is highly dynamic and subject to structural alterations. The high number of SNVs and CNVs found in such a small genome supports the importance of structural variations in shaping the grapevine genomes. These findings, even considering the small number of samples studied, may represent an important step forward for the identification of candidate genes for some of the most desired traits in breeding programs. As a future perspective, additional studies focusing on CNV genes should be performed on a bigger survey to verify the association to specific traits and to validate at a functional level the implication of candidate genes to precise phenotypes.

CONCLUSIONS

Taken together, our data demonstrate that plastic regions compose more than 26% of the grapevine genome and 8% is variant among different varieties. As stated previously, structural variations in plants were considered to be part of the so-called 'dispensable genome' and not necessary for survival (Morgante *et al.*, 2007). Nevertheless, recent studies on their importance reveal that the distinction between core and dispensable genomes is not immutable, and structural variations could be considered as 'conditionally dispensable' (Marroni *et al.*, 2014). Our data represent a further step in favor of this hypothesis.

We developed an approach that combined HTS, array CGH, FISH and qPCR for plant genome studies to describe the genomic structure of multiple genomes at the same time. For the *V. vinifera* L. species, this represents a landmark for future comparative studies.

EXPERIMENTAL PROCEDURES

Plant material and sequencing

We selected for sequencing four grape cultivars (cv): AR; It; RG; and TS, from a grape collection grown in the experimental field of the Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria (CREA)-Research Unit for viticulture and enology in Southern Italy (CREA-UTV).

The main features of the chosen varieties are detailed in Figure S6. Pedigree information for It, RG and AR were collected, while TS is an ancient variety of uncertain origin (Figure S7).

Total genomic DNA was isolated from young leaves using DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. The DNA quality and quantity was assessed by both gel electrophoresis (0.8% agarose) and spectrophotometer at 260 nm. The genomic DNA was used for preparing 76-bp paired-end libraries sequenced using the Illumina GAIIx platform.

The sequence data have been submitted to the Sequence Read Archive, under the study ID SRP009057 (https://wiki.nci.nih.gov/display/TCGA/Short+Read+Archive).

The genome assembly of the PN40024 inbred line (Jaillon *et al.*, 2007) was used as the reference to align the reads and 'illuminazed' as already reported for other genomes (Alkan *et al.*, 2009; Sudmant *et al.*, 2010; Mills *et al.*, 2011; Ventura *et al.*, 2011; Bickhart *et al.*, 2012; Prüfer *et al.*, 2012; Scally *et al.*, 2012; Prado-Martinez *et al.*, 2013; Freedman *et al.*, 2014; Montague *et al.*, 2014; Tamazian *et al.*, 2014). In addition, as the Illumina versus Sanger sequencing could differ in GC-rich regions as Illumina coverage decreases in higher GC regions, we corrected for the GC biases for each sequencing experiment separately, therefore minimizing any GC-dependent differences between Illumina and Sanger. The overall sequencing results were reported in Table S11. The *V. vinifera* L. chromosome, mRNA and peptide sequences were obtained from the grapevine genome project repository (GENOSCOPE) web site (http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/annotation/). We also downloaded *V. vinifera* L. whole-genome sequence reads and related sequence quality data (clip files) from the NCBI Trace Archive (ftp://ftp.ncbi.nih.gov/pub/TraceDB/vitis_vinifera/).

SNV discovery

To discover SNPs and small indels, we pooled all obtained sequences to reach a final coverage of 67.34 × (calculated as the summary of the coverage obtained for each variety) in order to increase the coverage of the analyzed genomes (Abecasis *et al.*, 2010) and aligned the reads generated from the genomes of the four table grape varieties to the reference genome assembly using BWA (default parameters). After converting alignment files to BAM format and removing PCR duplicates with SAMtools, we used the GATK software to discover and genotype SNPs and indels. For this purpose, we followed the 'best practices' guidelines in the GATK documentation (McKenna *et al.*, 2010). We filtered out calls with genotype quality <40 and read-depth <3, obtaining 4 740 493 SNVs (4 478 098 SNPs and 262 395 indels) considered the most reliable.

SNP- and indel-obtained data were analyzed separately to compare the concordance of the detected features among the varieties, tagging each window by a specific ternary code. In

particular, for SNPs, we assigned to each call a code 0 if the variety had the same nt of the reference (homozygosis for the same allele of the reference genome); code 1 if the variety had a different nt with respect to PN40024 in just a single allele (presence of a variant in heterozygosis); and code 2 if the variety had a different nt with respect to PN40024 on both the alleles (homozygosis for an allele absent in PN40024). For indels, we instead assigned to each call a code 0 if the variety had no indel; code 1 if the variety had both the indel and the reference allele (presence of a variant in heterozygosis); and code 2 if the variety had both alleles with an indel (homozygosis for a status different from that of PN40024). Variant calls in heterozygous condition, i.e. having code 1, could be considered either concordant or discordant with one of the two possible homozygous genomes. Thus, we checked for unique and common variants considering the heterozygous condition once concordant with the PN40024 status (code 0) and once concordant with the alternative homozygous status (code 2).

Thus, we checked for unique and common variants considering the heterozygous condition concordant with the PN40024 status and concordant with the alternative homozygous status. Moreover, we compared our SNP calls with previously published data (Di Genova *et al.*, 2014).

SNV validation

We selected 87 SNPs distributed on all chromosomes for validation by visually inspecting data with the Integrative Genomics Viewer tool, preferring those altering the coding sequence length of the genes but excluding those previously validated by Di Genova *et al.* (2014). We focused on SNPs that either produce or remove stop codons (gain or loss) in genes involved in metabolic processes.

For each SNP, PCR amplification was performed for all four varieties and the PN40024 reference. Amplification was carried out in 25 μL reactions with 1 × PCR buffer, 1.5 mM magnesium chloride, 200 μM dNTPs, 0.5 μM forward and reverse primers, 50 ng of DNA, and 0.03 U μL^{-1} Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, California, USA). The reaction was then cycled with the following conditions: initial denaturation at 95°C for 4 min, then 35 cycles at 95°C for 30 sec, 63–65°C for 40 sec, and 72°C for 40 sec; final extension was at 72°C for 5 min. PCR products were purified using QIAquick PCR Purification Kit (Qiagen). Samples were sequenced using the Sanger method and all variants were manually called by visual inspection.

Read-depth analysis and WSSD

We defined the SD content and estimated the absolute CN counts in the four genomes using a version of the WSSD approach (Bailey *et al.*, 2002) modified for HTS data (Alkan *et al.*, 2009). This algorithm leads to the detection of duplicated genomic regions, highlighted by a local excess of depth of coverage. We first masked common repeats as detected by RepeatMasker (http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/assembly/goldenpath/masked/) and simple tandem repeats smaller than 12 nucleotides detected by the Tandem Repeat Finder (Benson, 1999). We then aligned the Illumina reads requiring 95% sequence identity (equivalent to 5% sequence divergence) using the mrFAST aligner (Alkan *et al.*, 2009). Next, we calculated the absolute CN of non-overlapping windows of 1 kbp unmasked sequence (KbUS) using mrCaNaVaR version 0.31, and duplicated/deleted segments were predicted based upon excess depth-of-coverage in 5 KbUS sliding windows (Alkan *et al.*, 2009). Finally, we identified SDs as regions with at least five consecutive windows having a CN > 2.5. Similarly, we characterized regions with low read-depth of coverage (CN 1.5 and below) as deletions as previously reported (Alkan *et al.*, 2009, 2011).

CN comparison, digital CGH and CNVRs

Data from the above-mentioned read-depth analysis were parsed and analyzed to compare among the sequenced grape varieties the CN duplication/deletion status of each 1 KbUS window. The comparison was performed considering either the absolute CN state or the L2R.

In order to detect CN variations among the four sequenced genomes, we used an '*in silico* digital CGH' approach by modifying an algorithm previously described by Sudmant *et al.* (2010). In particular, the estimated CN of each 1 KbUS window for each variety was compared with the CN of the same window in PN40024. The L2R of that comparison was also calculated for each window. Similar to array CGH, this allowed for the detection of regions of gain or loss in each genome compared with the reference.

In order to detect large regions with CNVRs and statistically significant aberrations, we searched for regions >10 kbp showing gain or loss using a threshold of L2R >0.25 for amplifications and L2R <−0.25 for deletions. The detected CNVRs were then inspected in order to define aberrations that were variety-specific or common to a subset of grape varieties, using a ternary code similar to the procedures used to tag the SNVs. We checked for deletion, standard copies and amplification in PN40024, attributing to the regions 0, 1 and 2, respectively. The remaining digit of the code was instead assigned to the region with reference to the L2R status, i.e. 0, 1 and 2 were the tags for L2R values <−0.25, comprised between −0.25 and +0.25, and >0.25, respectively. With this rule, the code 901212 indicates that in the reference genome the region was amplified (2 as last digit), and that almost the same level of amplification was detected in the other two varieties (1), while a single variety shows a greater amplification (2) and a different variety shows a lower amplification (0). Finally, we also pairwise compared varieties by L2R calculation using the same procedures (Appendix S1-§*1 CNV calling*).

CNVR validation by array CGH, FISH assays and qPCR

We performed array CGH to confirm individual-specific and shared aberrations. Starting from the assembled *V. vinifera* L. genome sequence and using the online tool eArray provided by Agilent Technologies S.p.A., we designed a custom array containing 172 659 60-bp oligos with an approximate density of one probe every 2.8 kbp.

Total genomic DNA isolated from each of the four table grape cultivars was hybridized according to the manufacturer protocol (Version 7.1-Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis, Agilent Technologies S.p.A.) against the total genomic DNA of PN40024. PN40024 buds were provided by INRA-CNRGV [The French Plant Genomic Resource Center (http://cnrgv.toulouse.inra.fr)], Genoscope (http://www.genoscope.cns.fr/spip/spip.php?lang=en) and Unité de Recherche en Génomique Végétale [URGV – Plant Genomics Research (http://www-urgv.versailles.inra.fr/)].

Specific *V. vinifera* L. C0t-1 was prepared (Zwick *et al.*, 1997) from *V. vinifera* L. genomic DNA extracted from leaves (Crespan *et al.*, 1999) and used in each experiment.

All array CGH experiments were performed with a standard replicate dye-swap experimental design (reverse labeling of the test and reference samples).

To analyze the array results, we used the DNA Workbench Software (Lite Edition 6.5, Agilent Technologies S.p.A.), and followed

the instructions to normalize the signals and preprocess the data. Aberration calls were performed using the ADM-2 algorithm provided within the software (Appendix S1-§*1.2 Digital CGH vs Array CGH*).

The selection of the 43 BAC was focused on regions that were not validated using array CGH, and signal patterns were compared with CN predictions of all the 1 KbUS windows composing the genomic sequence contained in the clone (Table S12). For each experiment, we annotated the number (one to more than five) and intensity (single or tandemly duplicated) of signals in each of the four grape varieties. We classified a pattern as 'not assigned' when the CN could not be estimated. Moreover, we defined the validation as 'not inferable' when it was not possible to compare estimated CN with FISH interphase signals. Interphase nuclei were obtained using a previously described drop-spreading technique (Giannuzzi *et al.*, 2011).

FISH probes were derived from the *V. vinifera* L. PN40024 BAC library, which was developed by INRA-CNRGV.

BAC probes were directly labeled with Cy3-dUTP by nick-translation. Slide treatment and FISH hybridizations were performed as previously described (Giannuzzi *et al.*, 2011), and *V. vinifera* L. C0t-1 was used in each experiment. High-stringency, post-hybridization washes were made: 3 × 5 min at 60°C in 0.1 × SSC (1 × SSC = 150 mM NaCl, 15 mM sodium citrate, pH 7.0).

Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a cooled CCD camera, and 60 images were acquired for each experiment to confidently assign the CN.

Quantitative PCR validation experiments were performed, and 21 loci were selected to validate the CN predictions. Primers were designed using Primer3 software (http://primer3.ut.ee/; Appendix S1-§2. *Validation of CN estimation by qPCR*). The specificity of each primer pair was first tested in triplicate on the analyzed grapevine varieties, whose genomic DNA was previously extracted with the DNeasy Plant Mini Kit following manufacturer's instructions (Qiagen).

Each qPCR was performed in a final volume of 15 μL containing 50 ng DNA, 1.5 μL forward and reverse primers (1 μM), and 7.5 μL Brilliant III SYBR® MM (from Agilent Technologies S.p.A.).

Quantitative PCR was conducted in a LightCycler 480 instrument (Roche Applied Science, Penzberg-Germany). The LightCycler protocol began with an initial incubation step (50°C for 2 min), followed by the polymerase activation step (95°C for 10 min). After that, 40 cycles with a denaturing temperature of 95°C for 10 sec and the annealing and extension step at 60°C for 30 sec were set. After PCR amplification, the PCR products were completely denatured at 95°C for 15 sec, cooled to 55°C at a thermal transition rate of 4.4°C per second, and then heated to 95°C at a thermal transition rate of 2.2°C per second with continuous fluorescence monitoring in the SYBR Green channel (melting curve analysis). LightCycler 480 software (Roche Applied Science) was used to analyze the data.

To infer the CN of the investigated region, we used the relative standard curve method comparing with an endogenous reference gene arbitrarily taken as constantly diploid, the fructose-6-phosphate-2-kinase (primer forward 5′-TCTAAACCGGGTCCTC ACTG-3′ and primer reverse 5′-CCCGAGACTCAAGAACCTCA-3′; Appendix S1-§*3.3 Quantitative PCR assays*) as already reported (Muñoz-Amatriaín *et al.*, 2013).

Functional analysis

We classified the CNVRs based on common or unique phenotypes of the four varieties. Using the gene list available at the public site of the CRIBI Biotechnology Center of University of Padua (http://genomes.cribi.unipd.it/DATA/) and downloaded from the Grape Genome Browser, we selected all genes that map at least partially inside the polymorphic regions found by the *in silico* analysis and thus potentially being CN polymorphic genes. Then gene content and relative annotation were reported for each class and for each identified CNVR. The complete list of *V. vinifera* L. gene annotations and their correspondence with Grape Genome Browser annotation was downloaded from the VitisNet: Grapevine Molecular Networks database (http://www.sdstate.edu/ps/research/vitis/pathways.cfm; Grimplet *et al.*, 2009, 2012).

ACKNOWLEDGEMENTS

The authors thank Tonia Brown of the University of Washington (Seattle) for proof-reading the manuscript, and INRA-CNRGV [The French Plant Genomic Resource Center (http://cnrgv. toulouse.inra.fr)] for providing buds derived from PN40024 inbred line. This study was supported by a grant from the Apulia Region (PO FESR-FSE 2007-20013-Project TEGUVA cod.61/09) and the Italian Ministry of University and Research-MIUR (PON ''R&C''-2007-2013-Project ONEV- cod.00134/2011; PON02_00186_2866121 "ECO_P4"; and Project "Futuro in ricerca" 2010 RBFR103CE3). The authors have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

MFC, MV and DA designed the research and wrote the paper; MFC performed array CGH assays, analyzed the data and interpreted the results; PDA and CA performed the computational analyses; CRC, FA and GG performed FISH experiments; AM and CB performed SNV validation assays; CB and GC performed quantitative PCR assays; MFC, CB and RP performed functional analysis. All authors contributed to manuscript writing, reading and approved the final manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. CIRCOS diagram reporting WSSD results.
Figure S2. Comparison digital CGH versus WSSD.
Figure S3. Array CGH overall results.
Figure S4. FISH pattern for hyper-duplicated regions.
Figure S5. qPCR results: regression analysis.
Figure S6. Phenotypic features.
Figure S7. Pedigree information.
Table S1 SNP calls.
Table S2 SnpEff output
Table S3 SNP validation.
Table S4 WSSD_comparison
Table S5 Digital CGH results.
Table S6 Array CGH results.
Table S7 FISH results.
Table S8 Hyper-duplicated regions found by FISH assays.
Table S9 qPCR validation.
Table S10 List of the polymorphic subregions and genes.
Table S11 Sequencing results.
Table S12 BAC clones.
Appendix S1. Supporting methods and results.

REFERENCES

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A. and Consortium, G.P. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Alkan, C., Kidd, J.M., Marques-Bonet, T. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W. and Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Battilana, J., Emanuelli, F., Gambino, G., Gribaudo, I., Gasperi, F., Boss, P.K. and Grando, M.S. (2011) Functional effect of grapevine 1-deoxy-D-xylulose 5-phosphate synthase substitution K284N on Muscat flavour formation. *J. Exp. Bot.* **62**, 5497–5508.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.* **27**, 573–580.

Bickhart, D.M., Hou, Y., Schroeder, S.G. *et al.* (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Gen. Res.* **22**, 778–790.

Böttcher, C.D.C. (2012) Hormonal control of grape berry development and ripening. In *The Biochemistry of the Grape Berry* (Gerós, C.M., Delrot, H. and Sharjah, S., eds). Bentham Science, **Vol. 1**, pp. 194–217. Available at: http://ebooks.benthamscience.com/book/9781608053605/.

Cao, J., Schneeberger, K., Ossowski, S. *et al.* (2011a) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**, 956–963.

Cao, J., Shi, F., Liu, X., Jia, J., Zeng, J. and Huang, G. (2011b) Genome-wide identification and evolutionary analysis of Arabidopsis sm genes family. *J. Biomol. Struct. Dyn.* **28**, 535–544.

Chia, J.M., Song, C., Bradbury, P.J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Crespan, M., Botta, R. and Milani, N. (1999) Molecular characterisation of twenty seeded and seedless table grape cultivars (Vitis vinifera L). *Vitis*, **38**, 87–92.

Dal Santo, S., Vannozzi, A., Tornielli, G.B., Fasoli, M., Venturini, L., Pezzotti, M. and Zenoni, S. (2013) Genome-wide analysis of the expansin gene superfamily reveals grapevine-specific structural and functional characteristics. *PLoS ONE*, **8**, e62206.

DeBolt, S. (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol. Evol.* **2**, 441–453.

Di Genova, A., Almeida, A.M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., Pinto, M., Hinrichsen, P., Orellana, A. and Maass, A. (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* **14**, 7.

Doligez, A., Bertrand, Y., Farnos, M. *et al.* (2013) New stable QTLs for berry weight do not colocalize with QTLs for seed traits in cultivated grapevine (Vitis vinifera L.). *BMC Plant Biol.* **13**, 217.

Dong, Q.-H., Cao, X., Yang, G., Yu, H.-P., Nicholas, K.K., Wang, C. and Fang, J.-G. (2010) Discovery and characterization of SNPs in Vitis vinifera and genetic assessment of some grapevine cultivars. *Sci. Horticulturae*, **125**, 233–238.

Emanuelli, F., Battilana, J., Costantini, L., Le Cunff, L., Boursiquot, J.M., This, P. and Grando, M.S. (2010) A candidate gene association study on muscat flavor in grapevine (Vitis vinifera L.). *BMC Plant Biol.* **10**, 241.

Fasoli, M., Dal Santo, S., Zenoni, S. *et al.* (2012) The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell*, **24**, 3489–3505.

Freedman, A.H., Gronau, I., Schweizer, R.M. *et al.* (2014) Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet.* **10**, e1004016.

Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Curr. Opin. Plant Biol.* **9**, 436–442.

Fujita, Y., Yoshida, T. and Yamaguchi-Shinozaki, K. (2013) Pivotal role of the AREB/ABF-SnRK2 pathway in ABRE-mediated transcription in response to osmotic stress in plants. *Physiol. Plant.* **147**, 15–27.

Giannuzzi, G., D'Addabbo, P., Gasparro, M., Martinelli, M., Carelli, F.N., Antonacci, D. and Ventura, M. (2011) Analysis of high-identity segmental duplications in the grapevine genome. *BMC Gen.* **12**, 436.

Goff, S.A. (2011) A unifying theory for general multigenic heterosis: energy efficiency, protein metabolism, and implications for molecular breeding. *New Phytol.* **189**, 923–937.

Goff, S.A. and Zhang, Q. (2013) Heterosis in elite hybrid rice: speculation on the genetic and biochemical mechanisms. *Curr. Opin. Plant Biol.* **16**, 221–227.

Grimplet, J., Cramer, G.R., Dickerson, J.A., Mathiason, K., Van Hemert, J. and Fennell, A.Y. (2009) VitisNet: 'Omics' integration through grapevine molecular networks. *PLoS ONE*, **4**, e8365.

Grimplet, J., Van Hemert, J., Carbonell-Bejerano, P., Díaz-Riquelme, J., Dickerson, J., Fennell, A., Pezzotti, M. and Martínez-Zapater, J.M. (2012) Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Res. Notes*, **5**, 213.

Hall, D.E., Robert, J.A., Keeling, C.I., Domanski, D., Quesada, A.L., Jancsik, S., Kuzyk, M.A., Hamberger, B., Borchers, C.H. and Bohlmann, J. (2011) An integrated genomic, proteomic and biochemical analysis of (+)-3-carene biosynthesis in Sitka spruce (Picea sitchensis) genotypes that are resistant or susceptible to white pine weevil. *Plant J.* **65**, 936–948.

Haun, W.J., Hyten, D.L., Xu, W.W. *et al.* (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* **155**, 645–655.

Hou, Y., Liu, G.E., Bickhart, D.M. *et al.* (2011) Genomic characteristics of cattle copy number variations. *BMC Gen.* **12**, 127.

Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S.A., Ware, D., Wing, R.A. and Stein, L. (2010) Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza. *Plant J.* **63**, 990–1003.

Jaillon, O., Aury, J.M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jung, C.J., Hur, Y.Y., Yu, H.J., Noh, J.H., Park, K.S. and Lee, H.J. (2014) Gibberellin application at pre-bloom in grapevines down-regulates the expressions of VvIAA9 and VvARF7, negative regulators of fruit set initiation, during parthenocarpic fruit development. *PLoS ONE*, **9**, e95634.

Keller, M. and Tarara, J.M. (2010) Warm spring temperatures induce persistent season-long changes in shoot development in grapevines. *Ann. Bot.* **106**, 131–141.

Lijavetzky, D., Cabezas, J.A., Ibáñez, A., Rodríguez, V. and Martínez-Zapater, J.M. (2007) High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. *BMC Gen.* **8**, 424.

Liu, G.E., Hou, Y., Zhu, B. *et al.* (2010) Analysis of copy number variations among diverse cattle breeds. *Gen. Res.* **20**, 693–703.

Malacarne, G., Perazzolli, M., Cestaro, A. *et al.* (2012) Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. *PLoS ONE*, **7**, e29762.

Marques-Bonet, T., Kidd, J.M., Ventura, M. *et al.* (2009) A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, **457**, 877–881.

Marroni, F., Pinosio, S. and Morgante, M. (2014) Structural variation and genome complexity: is dispensable really dispensable? *Curr. Opin. Plant Biol.* **18**, 31–36.

Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J. (2010) Functional annotation, genome organization and phylogeny of the grapevine (Vitis vinifera) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* **10**, 226.

Matus, J.T., Aquea, F. and Arce-Johnson, P. (2008) Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. *BMC Plant Biol.* **8**, 83.

May, B., Lange, B.M. and Wüst, M. (2013) Biosynthesis of sesquiterpenes in grape berry exocarp of Vitis vinifera L.: evidence for a transport of farnesyl diphosphate precursors from plastids to the cytosol. *Phytochemistry*, **95**, 135–144.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A. and Stupar, R.M. (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* **159**, 1295–1308.

McKenna, A., Hanna, M., Banks, E. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Gen. Res.* **20**, 1297–1303.

Mills, R.E., Walter, K., Stewart, C. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Montague, M.J., Li, G., Gandolfi, B. *et al.* (2014) Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl Acad. Sci. USA*, **111**, 17230–17235.

Morgante, M., De Paoli, E. and Radovic, S. (2007) Transposable elements and the plant pan-genomes. *Curr. Opin. Plant Biol.* **10**, 149–155.

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T. *et al.* (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Gen. Biol.* **14**, R58.

Muñoz-Bertomeu, J., Arrillaga, I., Ros, R. and Segura, J. (2006) Up-regulation of 1-deoxy-D-xylulose-5-phosphate synthase enhances production of essential oils in transgenic spike lavender. *Plant Physiol.* **142**, 890–900.

Muñoz-Espinoza, C., Di Genova, A., Correa, J., Silva, R., Maass, A., González-Agüero, M., Orellana, A. and Hinrichsen, P. (2016) Transcriptome profiling of grapevine seedless segregants during berry development reveals candidate genes associated with berry weight. *BMC Plant Biol.* **16**, 104.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M. *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.

Prüfer, K., Munch, K., Hellmann, I. *et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature*, **486**, 527–531.

Roach, C.R., Hall, D.E., Zerbe, P. and Bohlmann, J. (2014) Plasticity and evolution of (+)-3-carene synthase and (-)-sabinene synthase functions of a sitka spruce monoterpene synthase gene family associated with weevil resistance. *J. Biol. Chem.* **289**, 23859–23869.

Saintenac, C., Jiang, D. and Akhunov, E.D. (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Gen. Biol.* **12**, R88.

Scally, A., Dutheil, J.Y., Hillier, L.W. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.

Sharp, A.J., Locke, D.P., McGrath, S.D. *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88.

Sudmant, P.H., Kitzman, J.O., Antonacci, F. *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.

Tamazian, G., Simonov, S., Dobrynin, P. *et al.* (2014) Annotated features of domestic cat – Felis catus genome. *Gigascience*, **3**, 13.

Tettelin, H., Masignani, V., Cieslewicz, M.J. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.

Velasco, R., Zharkikh, A., Troggio, M. *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE*, **2**, e1326.

Ventura, M., Catacchio, C.R., Alkan, C. *et al.* (2011) Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Gen. Res.* **21**, 1640–1649.

Yu, P., Wang, C., Xu, Q., Feng, Y., Yuan, X., Yu, H., Wang, Y., Tang, S. and Wei, X. (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Gen.* **12**, 372.

Zheng, L.Y., Guo, X.S., He, B. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Gen. Biol.* **12**, R114.

Żmieńko, A., Samelak, A., Kozłowski, P. and Figlerowicz, M. (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* **127**, 1–18.

Zwick, M.S., Hanson, R.E., Islam-Faridi, M.N., Stelly, D.M., Wing, R.A., Price, H.J. and McKnight, T.D. (1997) A rapid procedure for the isolation of C0t-1 DNA from plants. *Genome*, **40**, 138–142.

