nature genetics
OPEN

The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution

The International Peach Genome Initiative*

Rosaceae is the most important fruit-producing clade, and its key commercially relevant genera (*Fragaria*, *Rosa*, *Rubus* and *Prunus*) show broadly diverse growth habits, fruit types and compact diploid genomes. Peach, a diploid *Prunus* species, is one of the best genetically characterized deciduous trees. Here we describe the high-quality genome sequence of peach obtained from a completely homozygous genotype. We obtained a complete chromosome-scale assembly using Sanger whole-genome shotgun methods. We predicted 27,852 protein-coding genes, as well as noncoding RNAs. We investigated the path of peach domestication through whole-genome resequencing of 14 *Prunus* accessions. The analyses suggest major genetic bottlenecks that have substantially shaped peach genome diversity. Furthermore, comparative analyses showed that peach has not undergone recent whole-genome duplication, and even though the ancestral triplicated blocks in peach are fragmentary compared to those in grape, all seven paleosets of paralogs from the putative paleoancestor are detectable.

Rosaceae includes species grown for their fruits (for example, peaches, apples and strawberries), lumber (black cherry) and ornamental value (roses). The family encompasses a wide variety of fruit types (pomes, drupes, achenes, hips, follicles and capsules) and plant growth habits (ranging from herbaceous to cane, bush and tree forms). The species that produce drupe fruits (peaches, apricots, almonds, plums and cherries) are important agricultural crops worldwide (for example, 20 million tons of peach are produced per year; FAOSTAT 2010, http://faostat.fao.org/), providing vitamins, minerals, fiber and antioxidant compounds for healthy diets. With an increasing need to improve the sustainability of our fruit and forest tree resources, a fundamental understanding of the biology and genetics of key tree species is important. As genetic and genomic resources, fruit trees are unique in that domestication and intensive breeding have captured the variant alleles of genes that control basic tree growth, fruit development and sustainability. Peach (*Prunus persica* (L.) Batsch), which has been bred and cultivated for more than 4,000 years[1], is a highly genetically characterized tree species whose genome is important for both fruit and forest tree research. In peach, genetic and genomic efforts have identified gene-containing intervals controlling a large number of important fruit traits (**Supplementary Note** and **Supplementary Table 1**).

Rosaceous genomes offer one of the best systems for the comparative study of genome evolution. The diploid species representatives of this family (strawberry, rose, raspberry and peach) have very small genomes of 200–300 Mb[2–4]; however, they show a broad diversity in growth habit. To better exploit this resource, the availability of whole-genome sequences of key diploid species is crucial.

RESULTS

Sequencing, assembly and map integration

We report here the high-quality whole-genome shotgun assembly of a double haploid genotype of the peach cv. Lovell (PLov2-2N; $2n = 2x = 16$) with an estimated genome size of 265 Mb[5]. Using Arachne v.20071016 (ref. 6), we assembled a total of 3,729,679 Sanger sequence reads (8.47-fold final sequence coverage) in 391 major scaffolds (>1 kb) covering 226.6 Mb. We screened these scaffolds (**Supplementary Note**) and checked them for putative misassemblies; 234 were retained and 40 were anchored using 827 markers from an updated version of the previously published *Prunus* reference map[7] to form the final release of 224.6 Mb of the peach genome (Peach v1.0) organized in eight pseudomolecules (215.9 Mb, 96.1% of the total assembly) and 194 unmapped scaffolds with scaffold and contig N50/L50 values of 4 Mb/26.8 Mb and 294 kb/214.2 kb, respectively. Aligning a set of 13 finished fosmid sequences to the genome revealed 187 non–gap adjacent mismatches out of 442,732 bp aligned, yielding an average base-pair accuracy of 99.96%. We estimated the completeness of the euchromatic portion of the assembly by aligning 74,606 *Prunus* ESTs obtained from GenBank onto the assembly, and approximately 1% of the ESTs were not found (**Supplementary Note**, **Supplementary Tables 2–5** and **Supplementary Figs. 1–4**). An extensive check of the current release revealed a few misassembly and orientation issues that will be dealt with in an upcoming release (**Supplementary Note**, **Supplementary Fig. 5** and **Supplementary Tables 6–10**). We compared the quality of the peach sequence assembly to that of other plant genomes using the standards established in a previous publication[8] and noted a high level of contiguity and fraction of mapped sequences (**Supplementary Table 11**). We assigned the approximate positions of

*A full list of authors and affiliations appears at the end of the paper.

Received 12 June 2012; accepted 22 February 2013; published online 24 March 2013; doi:10.1038/ng.2586

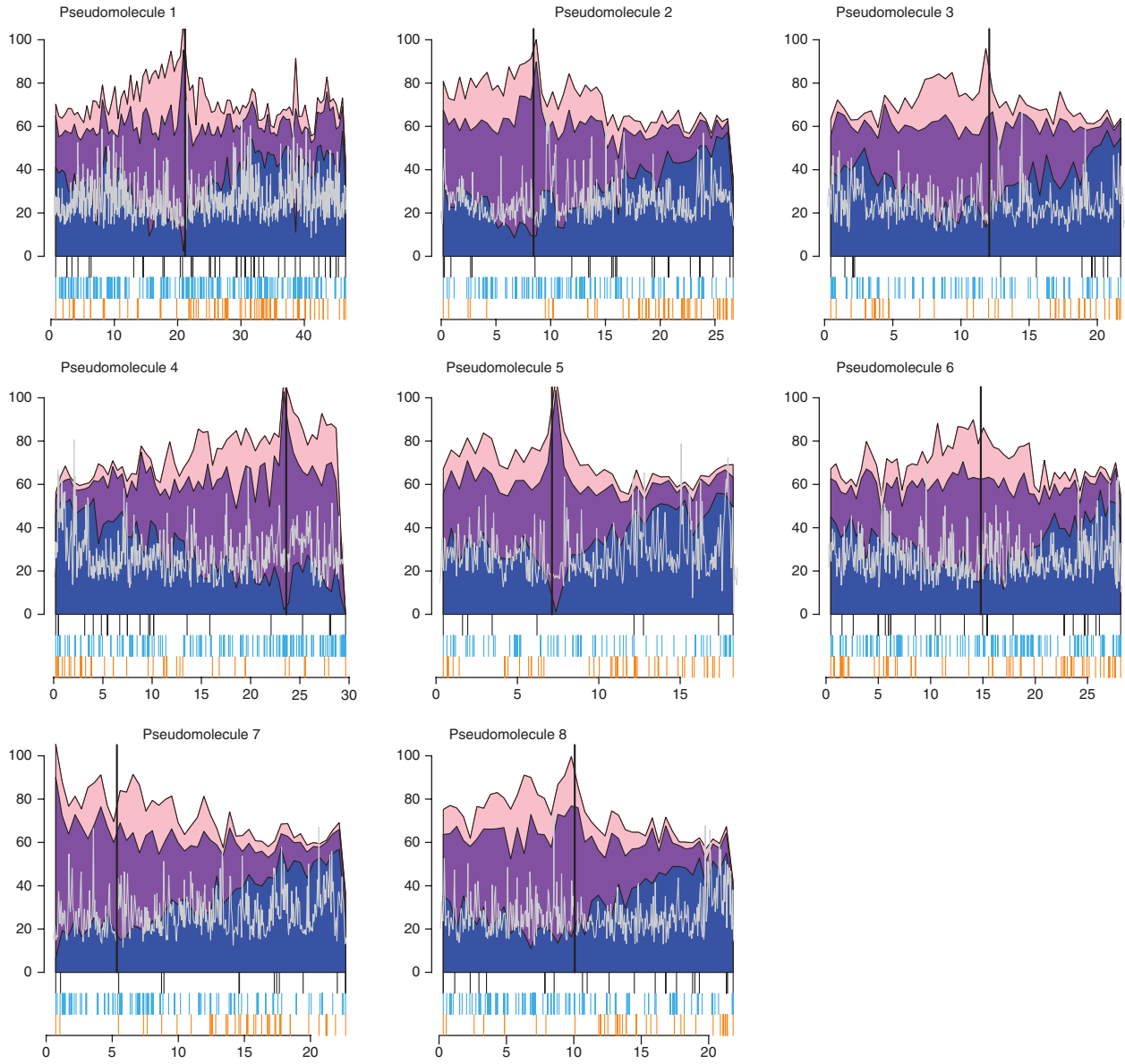

Figure 1 *P. persica* genome landscape. Plots for the eight pseudomolecules in the peach genome (scale along the *x* axis in Mb) showing the percentage of the genome (in nonoverlapping 500-kb windows) that consists of various annotated features, represented as stacked area graphs: type I transposable elements are shown in purple, type II DNA transposable elements are shown in pink, and genes are shown in blue. The gray line shows 100 times the mean r^2 value for all SNPs in 50-kb windows as an estimate of LD. The approximate position of each centromere is indicated with a vertical black bar. Below, three tracks of vertical lines show the positions of predicted miRNAs (black), noncoding RNAs (light blue) and tRNAs (orange).

the chromosome centromeric regions on the basis of gene-poor highly repetitive regions, with suppressed recombination observed in three different linkage maps (**Fig. 1** and **Supplementary Fig. 6**).

Repeat sequence annotation and gene prediction

Analysis of the repetitive fraction of the genome showed that long terminal repeat (LTR) retrotransposons comprise 18.56% of the genome, with Ty3-*gypsy* (9.97%) and Ty1-*copia* (8.6%) being represented in almost equal proportions. DNA transposons comprise 9.05% of the genome. Altogether, the identified transposable element sequences represent 29.60% of the genome, whereas 7.54% of the genome corresponded to uncharacterized repeats (**Supplementary Table 12**). These values are lower than those observed in apple (42.4%; ref. 9) and grape (44.5%; M. Morgante, unpublished data) but are higher than that observed in *Arabidopsis* (18.5%; TAIR 8, refs. 9,10), as could be expected on the basis of a proportionality of repeat content with genome size. The comparability of these estimates is also dependent on the methods used to identify repeats: the grape repeat identification procedure was very similar to that used in peach, making those two estimates highly comparable. Using a molecular paleontological approach[11], we estimated the insertion time of LTR elements on the basis of the nucleotide divergence of their LTRs. In the vast majority of cases, the LTR divergence was extremely low (**Supplementary Fig. 7**). Notably, 253 LTR retrotransposon elements (12.6% of the total) had identical LTRs. These data point to an extremely recent (possibly still ongoing) wave of retrotransposition for these elements. Analysis of the reverse transcriptase domains of both the Ty1-*copia* and Ty3-*gypsy* LTR retrotransposon subclasses identified in peach and woodland

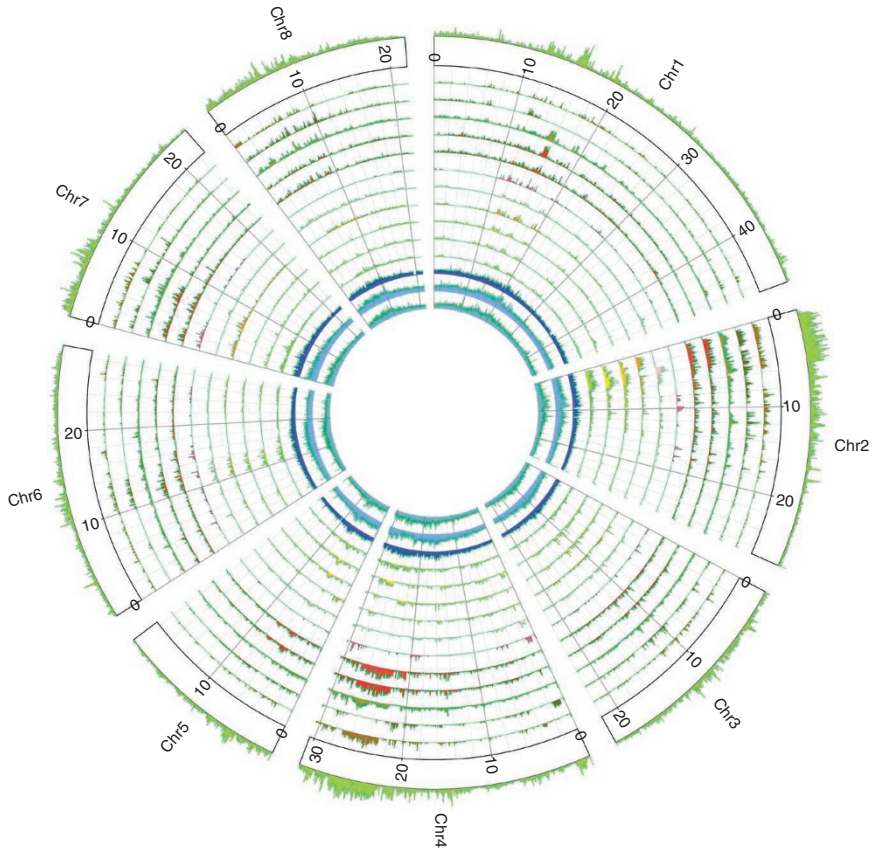

Figure 2 Nucleotide diversity distribution in peach. The outer track represents nucleotide diversity (π) in 50-kb nonoverlapping sliding windows estimated from a sample of 23 haploid genotypes (11 diploid accessions and the reference dihaploid Lovell). The 14 inner tracks depict the SNP frequency distributions for 50-kb nonoverlapping sliding windows in the ten peach accessions and four *Prunus* wild species compared to the reference individual (dihaploid Lovell). The order is (from the outside inward): *P. ferganensis* (E), Oro A (W), Shenzhou Mitao (E), Yumyeong (E), Sahua Hong Pantao (E), GF305 (W), Quetta (E), Earligold (W), IF7310828 (W), Bolero (W), F_1 Contender × Ambra (W), *P. kansuensis* (S), *P. davidiana* (S) and *P. mira* (S). E, eastern accessions; W, western accessions; S, wild species.

strawberry[2] showed that the elements from the two species are mixed within different clades (**Supplementary Figs. 8** and **9**), suggesting that the diversification into families largely predates the divergence of the two genera (**Supplementary Note**).

A total of 27,852 protein-coding genes and 28,689 protein-coding transcripts were predicted; of these, 24,423 have *Arabidopsis* homologs, 18,822 have Swiss-Prot homologs and 26,731 have TrEMBL homologs (**Supplementary Note**). The gene content in peach is considerably lower than those observed in apple (57,386; ref. 9) and poplar (45,654; ref. 12) but is similar to those in grape (30,434; ref. 13) and *Arabidopsis* (27,416; TAIR 10, ref. 10, http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp). The gene density in peach (1.22 genes per 10 kb on average) was higher than that in apple (0.78; ref. 9) but was lower than that in *Arabidopsis* (2.29; TAIR 10, ref. 10). In addition to protein-coding genes, we identified and annotated various noncoding RNA sequences. We integrated previously identified conserved microRNAs (miRNAs)[14] with the data presented in **Figure 1**. Furthermore, 474 transfer RNAs (tRNAs) decoding 20 amino acids, as well as 25 tRNA pseudogenes and 769 other noncoding RNAs, were predicted (**Supplementary Note**, **Supplementary Tables 13** and **14**).

Polyol biosynthesis and phenylpropanoid metabolism

To validate and improve the Peach v1.0 gene models, we manually annotated 672 gene models from 141 diverse gene families (**Supplementary Note**, **Supplementary Tables 15** and **16**) that are relevant to fruit quality–related traits. We considered the pathways for cell-wall metabolism, sugar metabolism and transport, abscisic acid and carotenoid synthesis, volatile compound metabolism, flavonoid and lignin biosynthesis, ethylene biosynthesis, MADS-box transcription factors and resistance genes. Only about one-third of the gene models required modification, of which approximately half were modified only in the UTR. Comparative and phylogenetic analyses carried out on the manually annotated gene families among peach and other sequenced species enabled the identification of members with specific roles in peach metabolic processes (for example, sorbitol metabolism and/or transport and aroma volatile compounds metabolism) and stressed common features with other Rosaceae species (**Supplementary Note**, **Supplementary Table 17**, **Supplementary Figs. 10** and **11**).

In Rosaceae, polyol biosynthesis[15] has a more prominent role than what is seen in other plant families. For example, in apple and peach, ~70% of translocated carbon is in the form of sorbitol[16,17]. Integrating the most recent Rosaceae molecular phylogeny[18] with data on sorbitol content[19], it is evident that leaf sorbitol synthesis and accumulation are restricted to the subfamily Spiraeoideae (for example, apple, peach and cherry), whereas in the subfamilies Rosoideae and Dryoideae, this polyol is comparatively absent[20,21]. Accordingly, in Spiraeoideae, a previous study[22] described sorbitol transporters (SOT) that substantially increase sorbitol uptake, and in cherry, two SOT-encoding genes are known to have a major role in sorbitol accumulation[23]. Besides those encoding transporters, other key genes in sorbitol metabolism encode A6PR (aldose 6-P reductase, which is rate limiting for sorbitol biosynthesis) and SDH (sorbitol dehydrogenase), which converts the alcohol into sugars in fruits[24,25]. We found that in contrast to other species with sequenced genomes, apple and peach *SDH* and *SOT* are large gene families (**Supplementary Table 17**). With the whole-genome sequence of peach (a diploid with no recent whole-genome duplication (WGD)), we were able to deduce from the position of these two species in the Rosaceae phylogeny that the specific gene family expansions probably occurred before the evolutionary split of the genera *Malus* and *Prunus*. Mapping of *SOT* gene clusters in peach and apple supports this conclusion (**Supplementary Note**). For *A6PR*, the expansion of gene number is evident only in apple. In strawberry (a Rosoideae species with low sorbitol content in leaves and fruit), the families *A6PR*, *SDH* and *SOT* do not have substantially different numbers of genes compared to other sequenced non-Spiraeoideae species (**Supplementary Table 17**). In conclusion, the genomic data support the species-specific massive polyol biosynthesis and accumulation as being linked, in part, to gene number expansion in particular gene families.

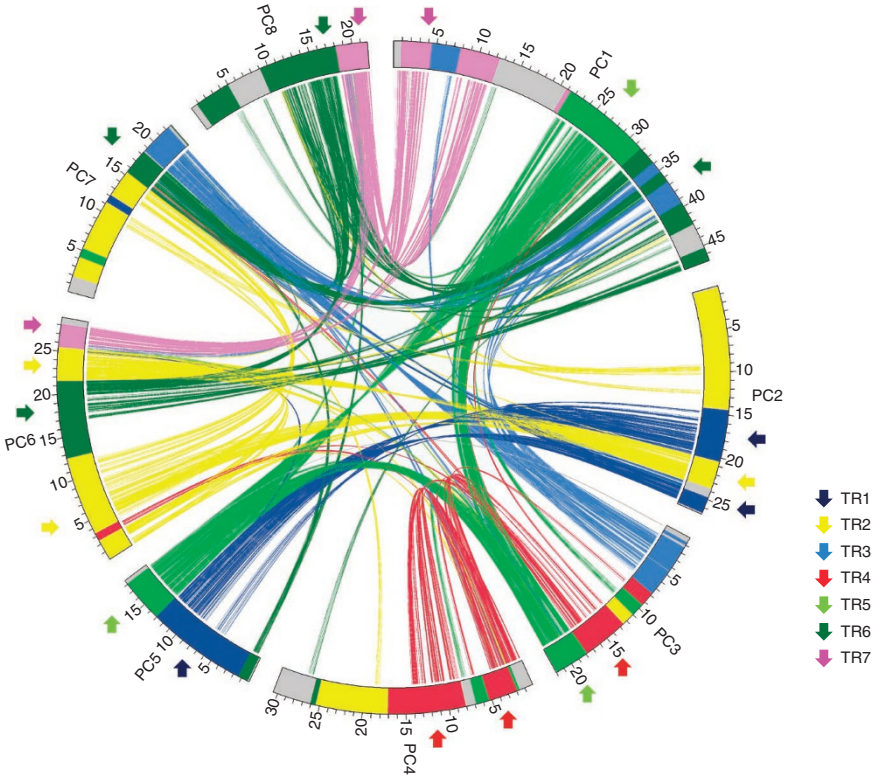

Figure 3 Duplicated and triplicated regions in the peach genome. Each line links duplicated regions in peach. The seven different colors represent each linkage group of the eudicot ancestor that existed before hexaploidization. Peach genomic regions are colored by their orthology to the grape genome. The lines are colored by the paralogous regions, and the order of precedence when paralogous regions have different ancestral origins is indicated by the colors of TR1, TR2, TR3, TR4, TR5, TR6, TR7 and gray. Seven major triplicated regions (TR1–TR7) are shown. PC, *Prunus* chromosome.

In addition to the unique aspects of polyol metabolism, phenylpropanoid metabolism in developing stone fruit is a unique biological system with spatially defined and well-timed switches in the flux of common precursors among the anthocyanin (early and late fruit development), lignin (stone formation) and free phenolic acids biosynthetic pathways. Sequence comparison with functional phenylpropanoid enzymes from other species identified 56 genes potentially involved in monolignol and anthocyanin biosynthesis in peach. This low number of genes specifying enzymes that produce this diversified set of secondary metabolic products is the most striking feature of the peach phenylpropanoid network. With two exceptions, p-coumarate 3-hydroxylase (C3H) and hydroxycinnamyl transferases (encoded by *HCT* and *HQT*), the peach phenylpropanoid toolbox has a minimal number of 'players' for a fruit-producing angiosperm tree. In peach, *C3H* (encoding the rate-limiting enzyme in monolignol biosynthesis) is represented by five members (four arranged in small tandem duplications on pseudomolecule 1 at ~42.6 Mb), four of which are notably expressed in fruit tissue, as confirmed by fruit ESTs and RNA sequencing (RNA-Seq) data. Similarly, the *HCT* and *HQT* gene family is also expanded (11 members in total) due to tandem duplication events in pseudomolecule 3 (10 members at 6–7 Mb). We confirmed expression for 7 out of 11 predicted *HCT* and *HQT* genes using the EST PASA alignments and/or RNA-Seq data. Thus, these two genes families encoding enzymes for crucial enzymatic reactions in monolignol and phenolic acids biosyntheses are expanded in peach compared to in the herbaceous Rosaceae species *Fragaria vesca*, which produces pseudocarp fruits, and poplar and grape, two woody plant perennial species with different reproductive biology. Only apple, which recently underwent a WGD[9], has a higher number of copies for the *C3H* gene family (**Supplementary Table 17**). Therefore, as was also evident in our analysis of peach polyol metabolism, the tandem gene duplication events in these two important gene families in phenylpropanoid metabolism are probably associated with specialization (that is, production of the lignified stone in *Prunus* fruits) and represent a *Prunus*-specific expansion of particular gene families in relation to specific phenotypic adaptations. We mapped the expansion and reduction of gene number for these and other families (**Supplementary Fig. 12**, which represents the phylogeny of Rosaceae adapted from a previous publication[18]).

Prunus diversity analysis and effects of domestication

To examine the genomic path of peach domestication, we resequenced 11 *P. persica* accessions (including the dihaploid Lovell PLov2-2N used for the reference assembly as a control) and one accession each of *Prunus ferganensis*, *Prunus kansuensis*, *Prunus davidiana* and *Prunus mira* (see the **Supplementary Note** and **Supplementary Table 18** for details about the resequenced accessions). Using a set of 953,357 high-quality SNPs identified in the peach and *P. ferganensis* accessions (the rationale for including *P. ferganensis* is discussed below), we estimated the nucleotide diversity for the eight pseudomolecules (**Fig. 2**, **Supplementary Tables 19**, **20** and **Supplementary Fig. 13**). The average nucleotide diversity (π) at the whole-genome level was 1.5×10^{-3}, with broad variation among individual pseudomolecules, where it ranged from 1.1×10^{-3} in pseudomolecule 1 to 2.2×10^{-3} in pseudomolecule 2. A markedly higher than average SNP diversity was evident at the top of pseudomolecule 2 and the bottom of pseudomolecule 4 (**Fig. 2**). The top of pseudomolecule 2 had a fivefold higher density of genes encoding the NB-LRR proteins compared to the rest of the genome. It also showed positive Tajima's D values, high amounts of haplotype diversity and linkage disequilibrium (LD) decay similar to the genomic average. As regions hosting resistance genes are known to evolve rapidly[26–28], this could explain the unusually high nucleotide diversity in this region. The bottom of chromosome 4 includes genes involved in fruit maturity time[29,30]. As peaches can be stored for only a few weeks, tens of varieties with overlapping maturity times must be available to the commercial market throughout the season (April to October in the northern hemisphere), and the varieties resequenced represent such a diverse set of maturity times. Breeding for this character then follows a divergent selection scheme that is compatible with the maintenance of a high amount of variability in this chromosomal region. In support of these hypotheses, we did not find wide differences in SNP diversity among regions in *P. kansuensis*, *P. davidiana* or *P. mira* (**Fig. 2**).

We grouped 11 *P. persica* accessions and one *P. ferganensis* accession according to their geographical origin (eastern compared to western). The rationale for including *P. ferganensis* in the group of

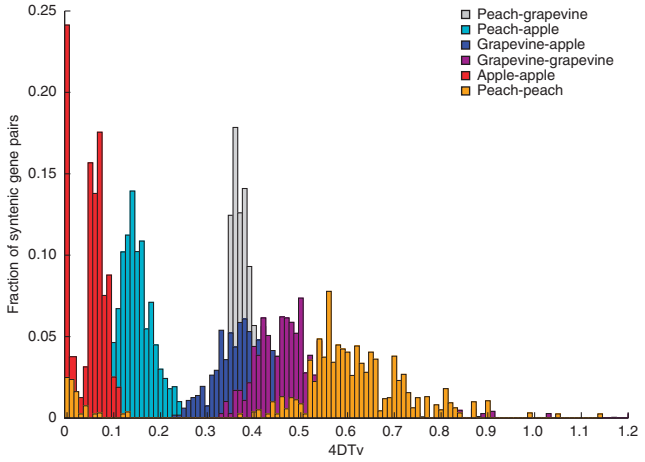

Figure 4 Distribution of 4DTv distance between syntenic gene pairs among peach, apple and grapevine. Segments of homologous genes were found by locating blocks of BLASTP hits with an E value of 1×10^{-10} or better with less than five intervening genes between such hits. The 4DTv distance between orthologous genes on these segments is shown.

P. persica accessions was that in phylogenetic analyses based on whole-genome SNPs (**Supplementary Fig. 14**), this wild species grouped with 'Shenzhou Mitao', a peach accession belonging to the northern China ecotypes that are most closely related to wild peaches[31]. This close relationship is supported by whole-genome sequence comparisons: *P. ferganensis* is indistinguishable from the cultivated varieties of peach (**Fig. 2**). Northwest China, between the Kunlun Shan mountains and the Tarim basin, is considered the center of origin of peach[1]. *P. ferganensis* comes from the Fergana Valley on the west side of the Tarim basin in central Asia. It shows some undomesticated traits, such as small fruit (70–80 g), absence of red coloration on the fruit skin, a typical pattern of unbranched leaf veins and a groove in the pit[32]. A plausible explanation for these results is that peach and *P. ferganensis* are in fact the same species, and *P. ferganensis* is a wild undomesticated peach or, more probably, represents an intermediate genome in peach domestication.

When we considered the molecular variation within the 12 accessions, there was a clear difference in nucleotide diversity: π was 1.6×10^{-3} for eastern varieties and 1.1×10^{-3} for western varieties (**Supplementary Table 20**). We evaluated the effects of the putative original domestication bottleneck that is supposed to have taken place in China 4,000–5,000 years ago by comparing the nucleotide diversity estimates obtained from a single accession of *P. davidiana*, a close interfertile wild relative of peach, with that of the domesticated Asian peach varieties, including *P. ferganensis*. The nucleotide diversity (π) was estimated at 4.8×10^{-3} for *P. davidiana*, the only outcrosser among the wild species analyzed, in contrast to the Asian peach varieties, which had a π value of 1.6×10^{-3}, highlighting the strong reduction of variability associated with domestication. The estimates of π obtained from just two haplotypes derived from a single individual may be subject to a high sampling variance but probably underestimate the diversity present within the species. A second bottleneck, related to the much more recent (16th–19th century) introduction of peach to the United States, is recorded by the difference in the π values between eastern and western varieties (1.6×10^{-3} compared to 1.1×10^{-3}, respectively). Further molecular evidence supporting recent historical bottlenecks of *P. persica* are a deficit of rare SNP variants (**Supplementary Fig. 15**), reflected in positive mean Tajima's D values (**Supplementary Fig. 16**), and LD r^2 values decaying rather slowly compared to those of other plant species[33,34] (the average r^2 value is reduced to one-half of its original value within 1 kb and to one-third in 10 kb) but with LD extending over hundreds of kilobases in specific chromosomal regions (**Fig. 1** and **Supplementary Fig. 17**). High local LD may result from selective sweeps related to domestication and breeding; quantitative trait loci for fruit size, a typical domestication trait, have been mapped in regions showing LD peaks on pseudomolecule 4 at ~2 Mb, ~8 Mb[35] and ~20 Mb[29,35,36] and on pseudomolecule 5 (15–17 Mb; **Fig. 1**)[35].

Comparative analysis and peach genome evolution

The radiation of eudicots started around 150 million years ago[37], and an accepted hypothesis maintains that the progenitor was hexaploid[38]. A corollary of the hypothesis is that the chromosomal state most similar to the paleohexaploid progenitor is present in extant members of the genus *Vitis*, where chromosomes are still assorted in triplets due to an unexpected maintenance of gene order along tens of millions of years[13]. Thus, comparing chromosomal segments of plant genomes to those of grape allows the description of gene and chromosomal events that have shaped the genomic state of living plant species.

We screened for duplicated regions (**Fig. 3**) in peach using DAGchainer[39], identifying a substantial number of duplicated regions. The data were largely sufficient to conclude that in peach the duplications were organized in seven major triplicated subgenomic regions (**Fig. 3** and **Supplementary Fig. 18a–g**). The dot plot analyses, however, indicated that the pattern of triplication was not as evident as that in grape (**Supplementary Fig. 19a,b**). This suggests that several interchromosomal rearrangements occurred during peach genome evolution. Regions without the paralogs corresponded to those with high SNP diversity (**Figs. 2** and **3**). We compared the grape and peach genomes using the Mercator program[40], which identifies segments with one-to-one orthology relationships across species rather than DNA regions having multiple syntenic partners. Notably, each grape segment, corresponding to part of one of the paralogous triplets of putative ancestor paleochromosomes, showed orthology to a single peach chromosome (**Supplementary Figs. 20** and **21**). This suggests that the homeologous subgenomes of grape and peach derive from the same paleohexaploid event that occurred before the emergence of Vitaceae and Rosaceae. In addition, the duplicated blocks in peach reside only in regions with the same prehexaploidy ancestral origin (**Fig. 3**)[38], suggesting that peach has not undergone recent WGD. Consistent with this argument, each paralogous region in peach is orthologous to one in grape[13] and two in poplar[12] (**Supplementary Fig. 22**). In summary, even though the triplicated blocks in peach are fragmentary compared to those in grape, all seven paleosets of paralogs are detectable.

To further analyze the evolutionary divergence of peach and other species, we calculated 4DTv (fourfold synonymous third-codon transversion)[12] rates (**Fig. 4**), which are indicative of the relative age of duplication. The 4DTv value peaked at 0.06 for paralog pairs in apple, highlighting the recent WGD in this species. A peak 4DTv value at 0.14 for the orthologs between peach and apple should correspond to species divergence. The orthologs between grape and peach or grape and apple showed 4DTv distances peaks at 0.36 and 0.38, respectively, which is consistent with the more ancient divergence between Vitaceae and Rosaceae. The 4DTv values between paralogs in peach and grape peaked at 0.56 and 0.50, respectively, again indicating that the hexaploidy in these eudicots occurred before the split of Vitaceae and Rosaceae (**Supplementary Note**).

DISCUSSION

The evolution of genes comprising the peach genome is intimately interwoven with the consequences of the exploitation of specific growth habits and tissue specialization (for example, drupe fruit development). For example, annotated gene members of 141 peach gene families identified and compared to those of six other fully sequenced diverse plant species (**Supplementary Tables 16** and **17**) are able to unravel unique evolutionary paths of important gene families, such as those involved in sorbitol metabolism and the phenylpropanoid pathway that leads to anthocyanin and lignin biosynthesis. The members of the phenylpropanoid gene network are the current subject of studies directed at examining the evolutionary changes in the phenylpropanoid gene toolbox that are associated with the expansion of angiosperm plants into different growth habit–related niches (herbaceous, cane bush or tree) representing the range of growth habit and fruit diversity in Rosaceae. This work, in conjunction with that of the recently published[41] comparison of the Peach v1.0 DNA sequence and the available genomes of apple[9] and strawberry[2], provides the substrate for developing an understanding of the changes in gene family repertoire and whole-genome sequence organization that are associated with fruit tree evolution.

The analysis with a resequencing approach using a range of cultivated accessions representing different germplasm pools and wild relatives allowed us to produce reliable estimates of nucleotide diversity and determine the effects of domestication and breeding-related processes on this diversity. The overall estimate of nucleotide diversity (π) was much lower than that recently obtained with similar approaches in *Medicago truncatula* (4.3×10^{-3}; ref. 33) and wild soybean (3.0×10^{-3}; ref. 42) but was similar to that in cultivated soybean (1.9×10^{-3}; ref. 42). Marked differences in diversity that were observed among chromosomal regions may have resulted from breeding activities and selection for specific traits, such as disease resistance and fruit maturity time. Having had available π data derived from accessions representing a substantial portion of the peach domestication and breeding history, we noted that they are consistent with those bottlenecks that have shaped the extant varieties of this crop. An original domestication bottleneck is supposed to have taken place in China 4,000–5,000 years ago, which was followed by the practice of vegetative propagation[1] and is reflected in the marked decrease in diversity observed between wild *P. davidiana* and the domesticated Asian peach varieties. After the dispersion of peach from China through Persia to Europe, a much more recent (16th–19th century) introduction of peach to the United States is represented by a few varieties that have subsequently served as the genetic foundation of the modern western breeding germplasm[43]. The effects of this second bottleneck are clearly reflected in the decrease of nucleotide diversity observed when moving from eastern to western varieties. These bottlenecks seem to have led to a considerable loss of diversity in western varieties in comparison to eastern varieties and wild relatives and have also resulted in a clear deficit of rare variants and a relatively slow LD decay.

Because of the resolution of mapped trait-containing intervals in peach for traits controlling fruit quality, fruit development and other important characteristics (**Supplementary Table 1**), the high-quality peach genome assembly, characterized by high contiguity, completeness and accuracy (**Supplementary Tables 4** and **11**), enables the rapid translation of genetic knowledge to actual gene members in specific gene families. Using comparative genomics approaches, this peach gene knowledge can be exploited for the improvement and sustainability of peach and other important tree species with less well-characterized genomes while at the same time enhancing our understanding of the basic biology of trees. In addition, this small and streamlined genome together with the absence of recent WGD makes it a key diploid tree genome that promises to provide important insights into terrestrial plant genome evolution. In this regard, its position within a family characterized by diploid species with very small genomes and an extreme diversity in plant growth habit offers an opportunity to investigate the specific gene and genome changes associated with the adoption of specific terrestrial growth niches.

URLs. The peach genome can be accessed using the genome browsers at http://www.rosaceae.org/species/prunus_persica/genome_v1.0, http://www.phytozome.net/peach and http://services.appliedgenomics.org/fgb2/iga/prunus_public/gbrowse/prunus_public/; RepeatMasker, http://www.repeatmasker.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AKXU00000000. The version described in this paper is the first version, AKXU01000000. Illumina Short reads have been deposited into NCBI Short Read Archive under accession number SRA053230. The 13 completely sequenced fosmids have been deposited into GenBank under accessions AC253537 to AC253549.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was jointly supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231 and the Ministero delle Politiche Agricole Alimentari e Forestali–Italy (MiPAAF, http://www.politicheagricole.it) through the project 'DRUPOMICS' (grant DM14999/7303/08). We would also like to thank the US Department of Agriculture (USDA) for their support of the peach genomics program through USDA National Institute of Food and Agriculture (NIFA) Specialty Crop Research Initiative (SCRI) grant 2010-2010-03255, the Robert and Louis Coker Chair for Plant Molecular Genetics for their grant to Clemson University, the Chilean government for supporting this work through FDI G02P1001 (Chilean Genome Initiative), Basal ProgramPB-16 and FONDAP CRG15090007, the Consolider-Ingenio 2010 Program (CSD2007-00036) from the Spanish Ministry of Science and Innovation, the French National Research Agency (ANR) for supporting this work through Chex-ABRIWG ANR/INRA 22000552, G. Reighard (Clemson University) for providing the leaf material of the Lovell double haploid, R. Quarta (Centro di Ricerca per la Frutticoltura di Roma (CRA-FRU)) and C. Pozzi (Fondazione Edmund Mach (FEM) S. Michele all'Adige) for their efforts in the drafting of the DRUPOMICS proposal, T. Pascal (Institut National de la Recherche Agronomique (INRA) Avignon) for providing leaf material for the resequencing, T. Candresse (INRA Bordeaux) for his critical reading of the manuscript, M. Troggio (FEM S. Michele all'Adige) for her help with the data analysis and G. Zhongshan (Zhejiang University) for the information provided about the Chinese peach accessions.

AUTHOR CONTRIBUTIONS

Please see the author list in the **Supplementary Information** for further details. I.V., A.G.A., F.S., J.S., B.S., M. Morgante and D.S.R. coordinated and managed the project (principal investigators). I.V., A.G.A., F.S., J.S., B.S., M. Morgante., D.S.R., S. Shu, S. Scalabrin and S.J. conceived and designed the experiments. J.S. (leader), D.S.R., M. Morgante, S.L., F.C., J.G., A.P., T.Z. and J.J. contributed to sequencing and assembly. I.V. (leader), J.S., M.T.D., P.A., S.T., A.G.A., T.Z., E.V., J.J., V.A., L.D. and S.F. contributed to linkage mapping, sequence anchoring and map integration. S. Scalabrin (leader), A.Z., M. Morgante, V.D., A.G.A., P.X., C.D.F., T.Z. and D.C. contributed to repeat analyses. S. Shu (leader), D.S.R., D.M., D.M.G. and S.P. contributed to gene annotation. A.G.A. (leader), F.S., H.S., G.V., C.W., A.O., P.T., D.S.H., D.M., I.V., F.C., T.Z., L.A.M., G.C., L.R., A.B., E.V., S. Scalabrin, A.S., C.D.F., S.G., R.F., J.M., E.M., S.F., B.L., D.B. and R.P. contributed to manual annotation, gene family characterization and small RNA analysis. M. Morgante (leader), I.V., F.S., R.T., F.M., S. Scalabrin, F.C., M.T.D., E.V., L.R. and M. Miculan contributed to resequencing and *Prunus* diversity analysis. S.J. (leader), D.M., D.S.R., F.S. and T.M.

contributed to comparative and peach genome evolution analyses. I.V., A.G.A., F.S., J.S., M. Morgante, D.S.R., P.A., S. Shu, S. Scalabrin and S.J. wrote the paper. I.V. and A.G.A. contributed equally.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Faust, M. & Timon, B. Origin and dissemination of peach. *Hortic. Rev. (Am. Soc. Hortic. Sci.)* **17**, 331–379 (1995).
2. Shulaev, V. *et al.* The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116 (2011).
3. Debener, T. & Linde, M. Exploring complex ornamental genomes: the rose as a model plant. *Crit. Rev. Plant Sci.* **28**, 267–280 (2009).
4. Meng, R. & Finn, C. Determining ploidy level and nuclear DNA content in *Rubus* by flow cytometry. *J. Am. Soc. Hortic. Sci.* **127**, 767–775 (2002).
5. Arumuganathan, K. & Earle, E.D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 208–218 (1991).
6. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
7. Howad, W. *et al.* Mapping with a few plants: using selective mapping for microsatellite saturation of the *Prunus* reference map. *Genetics* **171**, 1305–1309 (2005).
8. Chain, P.S.G. *et al.* Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).
9. Velasco, R. *et al.* The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
10. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
11. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
12. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
13. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
14. Barakat, A. *et al.* Genome wide identification and characterization of cold responsive microRNAs and siRNAs in *Prunus persica* by high-throughput sequencing. *BMC Genomics* **13**, 481 (2012).
15. Bieleski, R.L. Sugar alcohols. in *Plant carbohydrates I-Intracellular carbohydrates. Encyclopedia of Plant Physiology, New Series 13A.* (eds. Pirson, A. & Zimmermann, M.H.) 185–192 (Springer-Verlag, Berlin, 1982).
16. Moing, A., Carbonne, F., Zipperlin, B., Svanella, L. & Gaudillere, J.P. Phloem loading in peach: symplastic or apoplatic? *Physiol. Plant.* **101**, 489–496 (1997).
17. Klages, K., Donnison, H., Wünsche, J. & Boldingh, H. Diurnal changes in non-structural carbohydrates in leaves, phloem exudate and fruit in 'Braeburn' apple. *Aust. J. Plant Physiol.* **28**, 131–139 (2001).
18. Potter, D. *et al.* Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* **266**, 5–43 (2007).
19. Wallaart, R.A. M. Distribution of sorbitol in Rosaceae. *Phytochemistry* **19**, 2603–2610 (1980).
20. Fait, A. *et al.* Reconfiguration of the achene and receptacle metabolic networks during strawberry fruit development. *Plant Physiol.* **148**, 730–750 (2008).
21. Hancock, J.F. Strawberry. in *Temperate Fruit Crops in Warm Climates* (ed. Erez, A.) Ch. 17, 445–455 (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000).
22. Watari, J. *et al.* Identification of sorbitol transporters expressed in the phloem of apple source leaves. *Plant Cell Physiol.* **45**, 1032–1041 (2004).
23. Gao, Z. *et al.* Cloning, expression, and characterization of sorbitol transporters from developing sour cherry fruit and leaf sink tissues. *Plant Physiol.* **131**, 1566–1575 (2003).
24. Loescher, W.H. *et al.* Sorbitol metabolism and sink-source interconversions in developing apple leaves. *Plant Physiol.* **70**, 335–339 (1982).
25. Lo Bianco, R., Rieger, M. & Sung, S.S. Carbohydrate metabolism of vegetative and reproductive sinks in the late-maturing peach cultivar 'Encore'. *Tree Physiol.* **19**, 103–109 (1999).
26. Cannon, S.B. *et al.* Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *J. Mol. Evol.* **54**, 548–562 (2002).
27. McHale, L., Tan, X., Koehl, P. & Michelmore, R.W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7**, 212 (2006).
28. Xing, Y., Frei, U., Schejbel, B., Asp, T. & Lübberstedt, T. Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. *BMC Plant Biol.* **7**, 43 (2007).
29. Eduardo, I. *et al.* QTL analysis of fruit quality traits in two peach intraspecific populations and importance of maturity date pleiotropic effect. *Tree Genet. Genomes* **7**, 323–335 (2011).
30. Dirlewanger, E. *et al.* Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three *Prunus* species: peach, apricot and sweet cherry. *Heredity* **109**, 280–292 (2012).
31. Yoon, J. *et al.* Genetic diversity and ecogeographical phylogenetic relationships among peach and nectarine cultivars based on simple sequence repeat (SSR) markers. *J. Am. Soc. Hortic. Sci.* **131**, 513–521 (2006).
32. Okie, W.R. & Rieger, M. Inheritance of venation pattern in *Prunus ferganensis* × *persica* hybrids. *Acta Hortic.* **622**, 261–263 (2003).
33. Branca, A. *et al.* Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. USA* **108**, E864–E870 (2011).
34. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
35. Quilot, B. *et al.* QTL analysis of quality traits in an advanced backcross between *Prunus persica* cultivars and the wild relative species *P. davidiana*. *Theor. Appl. Genet.* **109**, 884–897 (2004).
36. Cantín, C.M. *et al.* Chilling injury susceptibility in an intra-specific peach [*Prunus persica* (L.) Batsch] progeny. *Postharvest Biol. Technol.* **58**, 79–87 (2010).
37. Chaw, S.M., Chang, C.C., Chen, H.L. & Li, W.H. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J. Mol. Evol.* **58**, 424–441 (2004).
38. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
39. Haas, B.J., Delcher, A.L., Wortman, J.R. & Salzberg, S.L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643–3646 (2004).
40. Dewey, C.N. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol.* **395**, 221–236 (2007).
41. Jung, S. *et al.* Whole genome comparisons of *Fragaria, Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies. *BMC Genomics* **13**, 129 (2012).
42. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
43. Scorza, R., Mehlenbacher, S.A. & Lightner, G.W. Inbreeding and coancestry of freestone peach cultivars of the eastern United States and implications for peach germplasm improvement. *J. Am. Soc. Hortic. Sci.* **110**, 547–552 (1985).

Ignazio Verde[1], Albert G Abbott[2,3], Simone Scalabrin[4], Sook Jung[5], Shengqiang Shu[6], Fabio Marroni[4,7], Tatyana Zhebentyayeva[2], Maria Teresa Dettori[1], Jane Grimwood[6,8], Federica Cattonaro[4], Andrea Zuccolo[4,9], Laura Rossini[10,11], Jerry Jenkins[6,8], Elisa Vendramin[1], Lee A Meisel[12,13], Veronique Decroocq[3], Bryon Sosinski[14], Simon Prochnik[6], Therese Mitros[15], Alberto Policriti[4,16], Guido Cipriani[1,7], Luca Dondini[17], Stephen Ficklin[5], David M Goodstein[6], Pengfei Xuan[18], Cristian Del Fabbro[4], Valeria Aramini[1], Dario Copetti[4], Susana Gonzalez[19], David S Horner[20], Rachele Falchi[7], Susan Lucas[6], Erica Mica[9], Jonathan Maldonado[21], Barbara Lazzari[10], Douglas Bielenberg[22], Raul Pirona[10], Mara Miculan[4], Abdelali Barakat[2], Raffaele Testolin[4,7], Alessandra Stella[10,23], Stefano Tartarini[17], Pietro Tonutti[9], Pere Arús[24], Ariel Orellana[19], Christina Wells[22], Dorrie Main[5], Giannina Vizzotto[7], Herman Silva[21], Francesco Salamini[10,25], Jeremy Schmutz[6,8], Michele Morgante[4,7] & Daniel S Rokhsar[6,26]

[1]Consiglio per la Ricerca e la Sperimentazione in Agricoltura (CRA)–Centro di Ricerca per la Frutticoltura, Rome, Italy. [2]Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, USA. [3]Institut National de la Recherche Agronomique (INRA), Université de Bordeaux, Unité Mixte de Recherche (UMR) 1332 Biologie du Fruit et Pathologie (BFP), BP81, Villenave d'Ornon Cedex, France. [4]Istituto di Genomica Applicata (IGA), Udine, Italy. [5]Department of

Horticulture and Landscape Architecture, Washington State University, Pullman, Washington, USA. [6]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [7]Dipartimento di Scienze Agrarie e Ambientali, University of Udine, Udine, Italy. [8]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, USA. [9]Scuola Superiore Sant' Anna (SSSA), Pisa, Italy. [10]Parco Tecnologico Padano, Lodi, Italy. [11]Dipartimento di Scienze Agrarie e Ambientali-Produzione, Territorio, Agroenergia, Università degli Studi di Milano, Milano, Italy. [12]Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, Santiago, Chile. [13]Centro de Biotecnología Vegetal Facultad de Ciencias Biológicas, Universidad Andrés Bello, Santiago, Chile. [14]Department of Horticultural Science, North Carolina State University, Raleigh, North Carolina, USA. [15]Energy Biosciences Institute University of California, Berkeley, California, USA. [16]Dipartimento di Matematica e Informatica, University of Udine, Udine, Italy. [17]Department of Fruit Tree and Woody Plant Sciences, University of Bologna, Bologna, Italy. [18]School of Computing, Clemson University, Clemson, South Carolina, USA. [19]Fondo de Investigación Avanzada en Áreas Prioritarias (FONDAP)–Center for Genome Regulation, Centro de Biotecnología Vegetal, Facultad de Ciencias Biológicas, Universidad Andres Bello, Santiago, Chile. [20]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Milano, Italy. [21]Laboratorio de Genómica Funcional y Bioinformática, Facultad de Ciencias Agronómicas, Universidad de Chile, La Pintana, Chile. [22]School of Agricultural, Forest and Environmental Sciences, Clemson University, Clemson, South Carolina, USA. [23]Istituto di Biologia e Biotecnologia Agraria, Consiglio Nazionale delle Ricerche (CNR), Lodi, Italy. [24]Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Centre de Recerca en Agrigenòmica Consejo Superior de Investigaciones Científicas (CSIC)-IRTA–Universitat Autònoma de Barcelona (UAB)–University of Barcelona (UB), Campus UAB, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain. [25]Istituto Agrario San Michele all'Adige (IASMA), Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige, Trento, Italy. [26]Center for Integrative Genomics University of California, Berkeley, California, USA. Correspondence should be addressed to I.V. (ignazio.verde@entecra.it), A.G.A. (aalbert@clemson.edu), M. Morgante (michele.morgante@uniud.it) or D.S.R. (dsrokhsar@gmail.com).

ONLINE METHODS

Genome sequencing, assembly and map integration. Sanger sequencing was used to generate paired-end reads from 2.8-kb, 4.4-kb, 7.8-kb, fosmid (35.3-kb) and BAC (69.5-kb) clones to generate 8.47× coverage (**Supplementary Table 2**). Sequence reads were assembled with Arachne v.20071016 (ref. 6) with the parameters maxcliq1 = 100, correct1_passes = 0 and BINGE_AND_PURGE = True. Scaffolds were aligned to a genetic map[7] to create pseudomolecules covering each chromosome (**Supplementary Fig. 5**). Markers were placed on the whole-genome shotgun (WGS) scaffolds using two methods. Simple sequence repeat (SSR)-based markers were placed using three successive rounds of electronic PCR (e-PCR)[44] with $N = 0$, $N = 1$ and $N = 3$. Markers that had a sequence associated with them, including RFLP and SNP markers, were best placed with BLAST and BLASTN.

Protein-coding gene annotation. Gene models were derived from weighted consensus prediction using several gene algorithms (FGENESH+ (ref. 45) and GenomeScan[46]), taking into account transcript assemblies (done with PASA[47]) and protein homology.

Repeats analysis. Both RepeatScout[48] and ReAS[49] were used to perform *de novo* identification of repeats. LTR retrotransposon structural identification was done using LTR_finder[50]. The REPET pipeline was used to generate a primary file of consensus transposable elements. Sequences were further manually curated using both BLASTX and Censor from the RepBase[51] database.

Transcriptome analysis. For RNA-Seq analyses, total RNA was extracted from different tissues: fruit at different ripening stages (cv. Imera, endocarp mesocarp and epicarp, S1-S2-S3-S4), roots (open pollinated cv. Yumyeong), fully expanded leaves (cv. Lovell), embryos and cotyledons (cv. Flavorcrest). In the case of fruit and seed, two bulks were prepared by pooling the RNA from different tissues and stages (**Supplementary Table 15**). The quality of the mRNA was tested using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA); the RNA Integrity Number (RIN) ranged from 7.3 for root to 9.8 for epicarp. RNA samples were processed using the RNA-Seq Sample Prep kit from Illumina (Illumina, Inc., CA, USA). Each library was loaded on one lane of an Illumina flowcell, and clusters were created using the Illumina Cluster Station (Illumina, Inc., CA, USA). Clusters were sequenced on a Genome Analyzer IIx (Illumina, Inc., CA, USA); 75-bp-long paired-end reads were obtained (**Supplementary Table 15**). Reads were initially preprocessed to remove possible contaminations from chloroplast, mitochondrion and ribosomal DNA (rDNA) and successively aligned to the peach reference genome using rNA[52] with default parameters.

Noncoding RNAs. tRNAscan-SE[53] was used to identify tRNA genes. In addition, we used INFERNAL 1.0.2 (ref. 54) with all RFAM[55] models that have previously generated hits from higher plants to identify small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and other noncoding RNAs (ncRNAs). Where predictions overlapped, the hit with the most significant *P* value was selected. Coordinates of all predictions were compared to gene predictions to identify intronic ncRNAs and ncRNAs that coincide with exonic predictions. ncRNA predictions within 5 kb of one another were considered clustered.

Resequencing and diversity analysis. For each accession, paired-end libraries were prepared as recommended by Illumina (Illumina Inc., San Diego, CA, USA) with minor modifications. Briefly, 1–3 µg of genomic DNA was sheared by nebulization, followed by standard blunt ending and 'A' addition. Then, Illumina adaptors were ligated to the ends of the fragments. After the ligation reaction and separation of unligated adaptors, samples were amplified by PCR to selectively enrich for those fragments in the library with adaptor molecules at both ends. The samples were quantified and quality tested using the NanoDrop ND-1000 UV-Vis Spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries were processed with the Illumina Cluster Generation Station following the manufacturer's recommendations and sequenced in one lane of the Illumina GA IIx or HiSeq2000 with 76 or 101 cycles per read. The CASAVA 1.7.0 version of the Illumina pipeline was used to process raw data.

Raw sequences were aligned against the Peach v1.0 reference genome (International Peach Genome Initiative, http://www.rosaceae.org/species/prunus_persica/genome_v1.0) after quality trimming using CLC Genomics Workbench 5.5 (CLC Bio, Aarhus, Denmark). Only reads that matched the reference sequence with ≥95% identity over ≥92% of their length and aligned to a single location were included in the alignment output file (reads that map equally well to multiple locations are not considered in the alignments). Average coverage was computed excluding zero-coverage regions. We only considered nucleotide positions in the reference that had a coverage ranging between 0.5 times and 1.5 times the average coverage and used a minimum minor allele frequency of 30% for identifying heterozygous polymorphisms. For SNP calling, the only bases used to support a SNP were those with quality score ≥20, and SNPs were only called if the 11 bp centered around the putative SNP had an average quality ≥15 and did not contain more than two SNPs and/or gaps.

We combined all SNPs identified in each accession to obtain a unique set for the ten *P. persica* accessions and the *P. ferganensis* accession that included only those variants that were present in nucleotide positions informative in at least four accessions. Nucleotide diversity (π) was computed among the 11 *P. persica* accessions (including the dihaploid PLov2-2N accession) and the *P. ferganensis* accession with internally developed scripts using the formula from Nei and Li[56]

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 \times \sum_{i=1}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

where x_i and x_j are the respective frequencies of the ith and jth sequences, π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences, and n is the number of sequences in the sample. The reference dihaploid PLov2-2N was considered as a haploid genotype, and for all other accessions, diploid genotypes were considered. The Watterson estimator θ_w was computed as K/a_n, where K is the number of segregating sites, and a_n is the $(n-1)$th harmonic number, with $n = 23$ being the number of haploid genotypes[57]. Tajima's D was obtained as the difference between the nucleotide diversity and Watterson estimator divided by the square root of the variance of that difference[58]. LD was estimated using r^2 in windows of size 50 kb using the same set of SNPs used to estimate nucleotide diversity and without any constraint on minor allele frequencies. Pairwise r^2 values for each pair of SNPs included in each window were estimated by maximum likelihood using the R package *genetics* (http://cran.r-project.org/web/packages/genetics/index.html). Decay of LD over distance was calculated according to the formula of Hill and Weir[59]. Background LD was estimated extracting one random SNP from each pseudomolecule and computing r^2 between SNPs. The procedure was iterated 50,000 times, and summary statistics were calculated. Average values of r^2 were calculated at every given distance from 1 bp to 50,000 bp and aggregated in bins of size 100 bp. If the sample size of a bin was lower than 1,000, larger bins were created. Analyses of linkage disequilibrium were performed using R[60].

Comparative analysis. Duplicated regions in the peach genome were analyzed using SynMap at the CoGe[61] website using BLASTZ[62] and DAGchainer[39] as the underlying software. Only syntenic groups including at least six gene pairs where the distance between two adjacent matches was lower than 200 kb were considered. The orthologous regions among species were detected using Mercator[40], which uses orthologous coding exons to define blocks of orthologous segments. The orthologous segments identified are those with one-to-one orthology relationships rather than any syntenic regions in which one region can have many syntenic regions. In finding orthologous segments, Mercator uses BLAT-similar anchor pairs in a modified k-way reciprocal best-hit algorithm[63]. In our analysis, two exons from each genome were selected as being similar if the BLAT score[64] of the pair was below 1×10^{-10}. The BLAT scores were computed using the translated protein sequences. The whole-genome sequence and annotation data of grape used in this analysis were downloaded from Genoscope (http://www.genoscope.cns.fr/). The whole-genome sequences of peach and grape were masked for repeats using the RepeatMasker program (see URLs), as well as the NMerge, WU-BLAST distribution and faSoftMask, Mercator[40] distribution utilities. Plots were obtained using Circos[65].

For the 4DTv calculations, orthologous genes were assigned within segments of syntenic genes. BLASTP was run between the two proteomes for comparisons with an E value cutoff of 1×10^{-10}. Segments were found by locating blocks of genes with a BLASTP hit with less than five intervening genes between such hits. The 4DTv is calculated as the number of transversions at all fourfold degenerate third codon positions divided by the number of fourfold degenerate third codon positions[12]. The 4DTv is corrected for multiple substitutions with the formula reported by Tang *et al.*[66].

44. Schuler, G.D. Sequence mapping by electronic PCR. *Genome Res.* **7**, 541–550 (1997).
45. Salamov, A.A. & Solovyev, V.V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
46. Yeh, R.F., Lim, L.P. & Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**, 803–816 (2001).
47. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
48. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
49. 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).
50. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
51. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
52. Vezzi, F., Del Fabbro, C., Tomescu, A.I. & Policriti, A. rNA: a fast and accurate short reads numerical aligner. *Bioinformatics* **28**, 123–124 (2012).
53. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
54. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
55. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33** (suppl. 1), D121–D124 (2005).
56. Nei, M. & Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269 (1979).
57. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
58. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
59. Hill, W.G. & Weir, B.S. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* **33**, 54–78 (1988).
60. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2012).
61. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant. Biol.* **1**, 181–190 (2008).
62. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
63. Hirsh, A.E. & Fraser, H.B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
64. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
65. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
66. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).