The genome of pear (*Pyrus bretschneideri* Rehd.)

Jun Wu, Zhiwen Wang, Zebin Shi, et al.

Genome Res. published online November 13, 2012
Access the most recent version at doi:10.1101/gr.144311.112

The genome of pear (*Pyrus bretschneideri* Rehd.)

Jun Wu[1,11], Zhiwen Wang[2,11], Zebin Shi[3,11], Shu Zhang[2,11], Ray Ming[4,11], Shilin Zhu[2,11,] M. Awais Khan[5], Shutian Tao[1], Schuyler S. Korban[5], Hao Wang[6], Nancy J. Chen[7], Takeshi Nishio[8], Xun Xu[2], Lin Cong[2], Kaijie Qi[1], Xiaosan Huang[1], Yingtao Wang[1], Xiang Zhao[2], Juyou Wu[1], Cao Deng[2], Caiyun Gou[2], Weili Zhou[2], Hao Yin[1], Gaihua Qin[1], Yuhui Sha[2], Ye Tao[2], Hui Chen[1], Yanan Yang[1], Yue Song[1], Dongliang Zhan[2], Juan Wang[2], Leiting Li[1,4], Meisong Dai[3], Chao Gu[1], Yuezhi Wang[3], Daihu Shi[2], Xiaowei Wang[2], Huping Zhang[1], Liang Zeng[2], Danman Zheng[5], Chunlei Wang[8], Maoshan Chen[2], Guangbiao Wang[2], Lin Xie[2], Valpuri Sovero[9], Shoufeng Sha[1], Wenjiang Huang[1], Shujun Zhang[3], Mingyue Zhang[1], Jiangmei Sun[1], Linlin Xu[1], Yuan Li[1], Xing Liu[1], Qingsong Li[1], Jiahui Shen[1], Junyi Wang[2], Robert E. Paull[7], Jeffrey L. Bennetzen[6], Jun Wang[2,10], Shaoling Zhang[1]

[1]Centre of Pear Engineering Technology Research, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China; [2]BGI-Shenzhen, Shenzhen 518083, China; [3]Institute of Horticulture, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021, China; [4]Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA; [5]Department of Natural Resources and Environmental Sciences, University of Illinois, Urbana, IL 61801, USA; [6]Department of Genetics, University of Georgia, Athens, GA 30602, USA; [7]Department of Tropical Plant and Soil Sciences, University of Hawaii, Honolulu, Hawaii 96822, USA; [8]Graduate School of Agricultural Science, Tohoku University, Aoba-ku, Sendai 981-8555, Japan; [9]Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA; [10]Department of Biology, University of Copenhagen, Copenhagen, Denmark; [11]These authors contributed equally to this work.

Co-corresponding authors. Shaoling Zhang (slzhang@njau.edu.cn) and Jun Wang (wangj@genomics.org.cn)

ABSTRACT

The draft genome of pear (*Pyrus bretschneideri*) using a combination of BAC-by-BAC and next generation sequencing is reported. A 512.0 Mb sequence corresponding to 97.1% of the estimated genome size of this highly heterozygous species is assembled with 194x coverage. High-density genetic maps comprising of 2,005 SNP markers anchored 75.5% of the sequence to all 17 chromosomes. The pear genome encodes 42,812 protein-coding genes, and of these, ~28.5% encode multiple isoforms. High quality of the assembly and annotation is assessed and confirmed using Sanger-derived BAC sequences along with transcriptome sequences of different tissues. Repetitive sequences of 271.9 Mb in length, accounting for 53.1% of the pear genome, are identified. Simulation of eudicots to the ancestor of Rosaceae has re-constructed nine ancestral chromosomes. Pear and apple have diverged from each other ~5.4 to 21.5 MYA, and a recent whole-genome duplication (WGD) event must have occurred 30-45 MYA prior to their divergence, but following divergence from strawberry. When compared with the apple genome sequence, size differences between apple and pear genomes are confirmed mainly due to presence of repetitive sequences predominantly contributed by transposable elements (TEs), while genic regions are similar in both species. Genes critical for self-incompatibility, lignified stone cells (a unique feature of pear fruit), sorbitol metabolism and volatile compounds of fruit have also been identified. Multiple candidate *SFB* genes appear as tandem repeats in the *S*-locus region of pear; while, lignin synthesis-related gene family expansion and highly expressed gene families of *HCT, C3'H, and CCOMT* contribute to high accumulation of both G-lignin and S-lignin. Expansion of *S6PDH, SDH*, and *SOT* along with evolutionary relationships of pear and apple have demonstrated their divergence from a common ancestor. Moreover, α-linolenic acid metabolism is a key pathway for aroma in pear fruit.

INTRODUCTION

Pear, the third most important temperate fruit species after grape and apple, belongs to the subfamily Pomoideae in the family Rosaceae. The majority of cultivated pears are functional diploids (2n = 34). As a popular fruit in the world market, pear has widespread cultivation on six continents with major production in China, USA, Italy, Argentina, and Spain (**Supplementary Fig. 1**). Pears are among the oldest of the world's fruit crops with more than 3,000 years of cultivation history (Lombard and Westwood, 1987), likely originating during the Tertiary period (65–55 MYA) in the mountainous regions of southwestern China, and from there spreading on to both the East and West (Rubtsov 1944; Zeven and Zhukovsky 1975). Central Asia and Eastern China are identified as two sub-centers of genetic diversity for pear (Vavilov 1951). The *Pyrus* genus is genetically diverse with thousands of cultivars, but it can be divided into two major groups, Occidental pears (European pears) and Oriental pears (Asiatic pears). At least 22 primary species are well-recognized in *Pyrus*; however, only a few species, including *P. bretschneideri*, *P. pyrifolia*, *P. ussuriensis*, *P. sinkiangensis,* and *P. communis* have been utilized for fruit production.

Herein, we report on a high-quality draft genome sequence of the diploid *P. bretschneideri* Rehd. cv. 'Dangshansuli' (also known as 'Suli'), the most important commercial Asiatic pear cultivar grown in the world (more than 4 million tons per year) and having more than 500 years of cultivated history in China. Pear is highly heterozygous due to self-incompatibility and interspecies compatibility. The genome is known to have an abundance of repetitive DNA sequences. In this study, a novel combination of BAC-by-BAC strategy, with Illumina sequencing technology, is used for the first time for *de novo* assembly of a highly heterozygous genome of this size with highly repetitive DNA sequences. This has demonstrated that a

complex plant genome sequence can be assembled and characterized using these technologies without availability of a physical reference. Additionally, we also report on primary factors contributing to genome size differences between pear and apple, both belonging to the subfamily Pomoideae, chromosomal evolution of Rosaceae, and on genes controlling valuable traits of pear, including self-incompatibility, lignified stone cells in flesh of fruit (unique to pear), sugar and aroma.

RESULTS AND DISCUSSION

Sequencing a highly heterozygous genome

The pear cultivar 'Suli' was first sequenced using a whole-genome shotgun (WGS) approach, but the quality of the assembled genome was poor. Analysis of 17-mer sequences revealed high levels of heterozygosity in the genome, and a 1-2% sequence divergence between alleles (**Supplementary Fig. 2**). To overcome this, a BAC-by-BAC strategy was used instead to sequence and assemble the pear genome. A total of 38,304 BACs was selected for sequencing, representing 7.6x genome equivalents. Two paired-end libraries with insert sizes of 250 bp and 500 bp, respectively, were constructed for each BAC, and sequenced at a combined 86x coverage using Illumina HiSeq 2000 (**Supplementary Table 1**). Each BAC was assembled individually prior to attempting whole genome assembly. In addition, WGS mate-pair libraries of 2Kb, 5Kb, 10Kb, 20Kb, and 40 Kb were constructed and sequenced at 24x coverage to build super-scaffolds; moreover, paired-end libraries of 180, 500, and 800 bp were constructed and sequenced at 83x coverage to fill in gaps (**Supplementary Table 2**). All BAC sequences were pooled for overlap-layout-consensus (OLC) assembly, identical sequences were merged, and redundant bases filtered out from overlapping lengths. The resulting contigs were assembled

into scaffolds by WGS paired-end reads of large-insert libraries (2 to 40 kb), and gaps were filled with WGS paired-end reads of small-insert libraries (180 to 800 bp).

Quality of the assembly was assessed by aligning scaffolds to five fully-assembled BAC sequences. The coverage ratios of BAC1, BAC2, BAC4, and BAC5 were over 98% with good synteny of scaffolds (**Supplementary Table 3, Supplementary Fig. 3**), while the coverage ratio of BAC3 was 90% as it had a 12K fragment that did not align to scaffold 227.0. This was attributed to differences between the two haplotypes. One haplotype was assembled in the final scaffold, but BAC3 belonged to an unassembled haplotype, even though both have been assembled in the BAC-to-BAC assembly step (**Supplementary Fig. 4**).

The assembled pear genome consists of 2,103 scaffolds with N50 at 540.8 kb, totaling 512.0 Mb with 194x coverage, close to the estimated size of 527 Mb (**Table 1**). Among 2,005 SNP markers in the genetic map, 100% of SNPs are anchored to 796 scaffolds, 386.7 Mb, representing about 75.5% of the assembled genome (**Supplementary Fig. 5**).

Heterozygozity features of the pear genome

A total of 3,402,159 reliable SNPs were identified in 'Suli'. Using the same filtering standard, 333,443,735 reliable genome bases were identified, thus the frequency of SNPs in this genome was about 1.02%. Heterozygosity of pear was higher than that of other plants, such as papaya (0.06%; Ming et al., 2008), pigeonpea (0.067%; Varshney et al., 2012), black cottonwood (0.26%; Tuskan et al., 2006), and date palm (0.46%; Al-Dous et al., 2011), but lower than that of grape (7%; Jaillon et al., 2007). The distribution profile of SNPs showed that 87.1% of SNPs were within 50 bp of each other, and nearly 50% were within less than 10 bp from an adjacent SNP (Supplementary Fig. 6). In contrast to frequency of SNPs within the whole genome, genes had lower frequencies of SNPs, of 0.84%, along with 0.70% for CDS, 0.95% for

introns, and 0.90% for UTRs. It was assumed that the frequency of SNPs in CDS was attributed to its conserved protein-coding function. A total of 26,249 genes had SNPs (Supplementary Fig.7). Of those, 13,794 genes had SNPs of less than 1%, and 4,346 genes had SNPs of more than 2%. These genes were enriched in major functional categories including protein kinase, disease resistance protein, cell division protein, ion transfer, and transcription factor. Genes with significantly high frequencies (>20%) of SNPs belonged to those with basic functions, including membrane, cell wall, cell division, and methylation, among others (Supplementary Fig. 8). Due to presence of SNPs, 1,300 genes changed from coding for amino acids to stop codons (nonsense mutations), and 500 genes changed from stop codons to other amino acids; these genes were enriched for biological function including cell division, protein kinase, and WD40 protein.

Influence of repetitive sequences on genome size variation

A combination of structure-based analysis and homology-based comparisons identified 271.9 Mb repetitive sequences, accounting for 53.1% of the current assembly of the pear genome (**Supplementary Table 4**). The most abundant transposon families were *Gypsy* and *Copia*, contributing for 25.5% and 16.9% of the genome, respectively (**Supplementary Table 5)**. Long Terminal Repeat (LTR) retrotransposons exhibited family-specific, non-uniform distributions along chromosomes. *Copia*-like elements were spread along the whole chromosome, including gene-rich euchromatic regions; whereas, *Gypsy*-like elements were overrepresented in gene-poor heterochromatic regions (**Fig. 1, Supplementary Fig. 9**). The most abundant DNA transposable elements (TEs) were PIF/*Harbinger* and hAT-Ac elements, representing 2.7% and 2.1% of nuclear DNA, respectively (**Supplementary Table 5)**. Although widely dispersed

throughout the genome, transposon-related sequences were most abundant in centromeric regions (**Fig. 1, Supplementary Fig. 9**).

Structural searches identified 645 reliable intact TEs and 19 Solo-LTRs in pear (**Supplementary Table 6**). These intact elements masked 34.0% of the assembly, accounting for ~70.0% of the total repeats. Of 299 intact LTR retrotransposons, 144 belonged to the *Copia* superfamily and 31 to the *Gypsy* superfamily. Low numbers of intact *Gypsy* elements did not suggest that *Gypsy* elements were relatively rare as large numbers of *Gypsy* RT domains were detected. By searching flanking sequences of DDE and RT domains, 10 intact hAT, 8 PIF/*Harbinger*, 2 CACTA, 38 LINEs, and 288 MITE elements (yielding 221 exemplars) were found (**Supplementary Table 7**). A partial reason for detecting relatively few intact TEs (357, excluding MITEs) was due to incomplete sequencing of individual BACs, thus leaving gaps in the assembly (especially in terminal repeats). This suggested that an element could not be deemed intact by any of the structure-based search algorithms used. However, these methods yielded ~3000 intact TEs in a ~500 Mb WGS assembly of the genome of the grass species *Setaria italica* (Bennetzen et al., 2012). Thus, it was likely that the pear genome yielded few intact elements in this current analysis either due to abundance of elements that were structurally rearranged or due to presence of more than a single copy of elements of the same families on any given BAC. This could be attributed to yield of high copy numbers of many families and/or insertion preferences during clustering, as noted for *Helitrons* in maize (Yang and Bennetzen, 2009).

Of a total of 603.9 Mb assembled data of the apple genome, an estimated 362.3 Mb repetitive sequences has been reported (Velasco et al. 2010). However, the non-repeat region of apple and pear is of almost equal size (241.6 Mb for apple and 240.2 Mb for pear). Thus,

the difference in repetitive sequences of assembled sequences of apple and pear is 90 Mb, mainly consisting of two forms of TEs, *Gypsy* and LINE **(Fig. 2)**. Additionally, a large portion of unassembled sequences in apple has been deemed as repetitive sequences (Velasco et al. 2010). Assembly of highly repetitive sequences is a major limitation for *de novo* sequencing of a heterozygous genome, such as pear, using WGS and next-generation sequencing technologies (Birney 2010). This is particularly true for TE families that have undergone recent amplifications. The BAC-by-BAC approach used in this study has ensured a relatively accurate assembly of TEs in the pear genome as TEs in different BACs would have rare effects during assembly of these BACs, although assembly of fully-intact elements will be rare when TEs either contain terminal repeats (e.g., LTR retrotransposons) or when more than a single copy of the same TE is found in a specific BAC. Based on these findings, observed genome size differences between apple and pear are mainly due to repetitive sequences predominantly contributed by TEs, while genic regions are similar in both species.

LTRs with complete structures in pear are predicted to estimate insertion time via distances between 5' and 3' solo-LTRs. These findings indicate that pear has a high LTR expansion rate; wherein, a recent two-fold increase in LTR numbers must have occurred, compared to other sequenced plant species (Arabidopsis 2000; Ming et al. 2008; Shulaev et al. 2010; Velasco et al. 2010). This suggests that the pear genome is in continuous expansion (**Supplementary Fig. 10**). However, these results may also be influenced by the method of assembly.

Gene annotation and transcriptome sequence analysis

By combining *ab initio* gene prediction and protein alignment prediction, 42,767 protein-coding genes were annotated. Comparisons of transcriptome sequences to gene models using Illumina RNA-Seq sequences provided empirical support for these predictions. This gene

prediction approach proved highly effective, as 23,843 (55.7%) hybrid gene models were supported by 25,365 (93.9% of 27,008 transcripts with complete Open Reading Frames (ORFs)) transcript-based sequences (**Supplementary Fig. 11**). After integrating and then adding novel transcriptome-based genes, a total of 58,596 transcripts constituted 42,812 gene loci, among which 12,217 (28.5%) genes encoded multiple isoforms. Thus, gene prediction based on whole genome assembly in pear was credible. On average, gene models consisted of transcript lengths of 2776 bp, coding lengths of 1172 bp, and means of 4.7 exons per gene, both similar to those observed in apple (Velasco et al. 2010) and Arabidopsis (Arabidopsis 2000). A total of 89.5% of gene models had matches in at least one of the public protein databases. These findings also confirmed completeness of the pear genome sequence coverage. In addition, 297 micro-RNAs (miRNAs), 1148 transfer RNAs (tRNAs), 697 ribosomal RNAs (rRNAs), and 395 small nuclear RNAs (snRNA) were identified in the pear genome (**Table 1**).

The number of genes found in pear is similar to that found in other sequenced plants of equivalent genome size, but much lower than that of the closely-related apple genome (Velasco et al. 2010). The pear genome has been sequenced using a BAC-by-BAC approach, resolving problems of assembling a heterozygous genome. In contrast, the apple genome has been sequenced using a WGS approach, wherein some alleles may have been annotated as individual genes. This is demonstrated by alignment of a single unique chromosome region to two overlapping scaffolds in apple (**Supplementary Fig. 12**). The assembly of two scaffolds for a single genomic region resulted in over-estimation of the assembled genome and gene numbers in apple. After filtering out of overlapping genes in apple chromosomes, the gene number in apple dropped from either 61,334 (based on NCBI) or 57,386 (based on the published report; Velasco et al. 2010) down to 45,293. This indicates that the numbers of genes in apple and

pear are almost equal.

The average gene density in pear is one gene per 12 kb, with genes being more abundant in sub-telomeric regions (**Fig. 1, Supplementary Fig. 9**), as previously observed in other sequenced plant genomes. Gene elements in pear, including lengths of miRNAs, distribution of CDS, and exons and introns are normally distributed when compared with those of five other plant species, including apple (*Malus* × *domestica*) (Velasco et al. 2010), strawberry (*Fragaria vesca*) (Shulaev et al. 2010), Arabidopsis (*Arabidopsis thaliana*) (Arabidopsis 2000), grape (*Vitis vinifera*) (Jaillon et al. 2007), and black cottonwood (*Populus trichocarpa*) (Tuskan et al. 2006) (**Supplementary Fig. 13**). Moreover, the pear genome maintains higher numbers of genes for transport and catalysis within the 'molecular function' gene ontology (GO) category; cellular process, protein metabolism, and biological regulation within the 'biological processes' GO category; and cell, intracellular, and membrane within the 'cellular component' GO category (**Supplementary Fig. 14)**.

WGD and divergence of *Pyrus*

A total of 13,372 pairs of paralogous genes in pear are aligned in 870 blocks (≥ 7 gene pairs per block). A four-fold degenerate site transversion (4dTv) of these blocks was calculated, and corrected by HKY. From the distribution of 4dTv (**Fig. 3**), it can be concluded that there are two significant groups of blocks, suggestive of WGD events in pear, including a recent event with 4dTv of ~0.08 and an ancient event with 4dTv of ~0.5. Both WGD events are shared by apple and pear, but strawberry does not share the recent WGD event. Distribution of 4dTv values suggests that divergence between pear and apple must have occurred after the recent WGD event. To estimate time of occurrence of these two duplication events, a total of 16,335 paralogous gene pairs within 5,593 gene families with substitutions per synonymous site (Ks)

values lower than two are selected. Based on Ks values in pear, the main peak ranges from 0.15 to 0.3, while the secondary peak ranges from 1.5 to 1.8, similar to that found in apple (**Supplementary Fig. 15 A and B**). As in apple, the recent WGD in pear must have occurred at 30-45 MYA (Velasco et al. 2010), while the ancient WGD must have resulted from an acknowledged paleohexaploidization event that took place ~140 MYA (Fawcett et al. 2009). The divergence time of eight sequenced plant species (Arabidopsis 2000; Jaillon et al. 2007; Khurana and Gaikwad 2005; Ming et al. 2008; Shulaev et al. 2010; Tuskan et al. 2006; Velasco et al. 2010), including pear, apple, strawberry, papaya, grape, black cottonwood (poplar), Arabidopsis, and rice, has been estimated according to known ranges of divergence time and the phylogenetic tree (**Supplementary Fig. 16 A**). It is estimated that pear and apple must have diverged from each other about 5.4 to 21.5 MYA (**Supplementary Fig. 16 B**).

Evolution of chromosomes in Rosaceae

Collinearity analysis between pear and two sequenced rosaceous species, apple and strawberry, has revealed that pear and apple share similar chromosome structures as well as organization **(Supplementary Fig. 17 A and B)**. All 17 chromosomes of pear displayed good homology with corresponding chromosomes of apple (**Supplementary Fig. 17 A)**. Based on self-collinearity of pear (Supplementary Fig. 17 B and Fig. 1 (A)), it is easy to identify good syntenic chromosome pairs in pear, such as LG3 and LG11, LG5 and LG10, LG9 and LG17, LG13 and LG16, as well as rearrangement of chromosomes as identified in the apple genome (Velasco et al. 2010). Most collinear regions between pear and strawberry reveal that one chromosome in strawberry corresponds to two chromosomes in pear **(Supplementary Fig. 17 C1 and C2**). For example, LG1 in strawberry corresponds to Chr2 and Chr15 in pear, and similarly LG2 to Chr5 and Chr10, LG3 to Chr3 and Chr11, LG4 to Chr13 and Chr16, as well as LG5 to Chr6 and Chr14, respectively.

It appears that LG2, LG3, LG5, and LG6 of strawberry are formed by fragmentation and recombination of ancestral chromosomes (**Fig. 4)**. In this study, ancestral chromosomes of paleohexaploid eudicots have been re-constructed to the ancestor of Rosaceae by collinearity between strawberry and grape genomes **(Supplementary Fig. 17 D)**. Using relationships of strawberry and Rosaceae and those of strawberry and eudicots, a simulated process of eudicots to Rosaceae has been developed. Results have revealed that triplication of seven ancestral chromosomes of eudicots may have undergone additional rearrangements, yielding nine ancestral chromosomes of Rosaceae (**Fig. 4**).

Disease resistance–related genes

A total of 396 nucleotide-binding site (NBS)-containing *R* genes were identified in pear (**Supplementary Table 10**). This was similar to that found in soybean (*Glycine max*) (392) and in poplar (402), and about 39.9% and 74.0% of that found in apple (992) and in rice (*Oryza sativa*) (535), respectively, but higher than that detected in Eurosids II, including both cacao (*Theobroma cacao*) (253) and *Arabidopsis* (178). However, the observed two-fold difference of numbers of *R* genes between pear and apple might have been over-estimated. Furthermore, pear *CC-NBS–LRR* genes outnumbered *TIR–NBS–LRRs*, which were similar to that observed in both grape and poplar, but in contrast to that found in apple, soybean, and *Arabidopsis*. In addition to *NBS* genes, the pear genome contained 403 *LRR-kinase* genes and 11 additional *CC-LRR-Kinase* genes, which is higher than that found in both apple (320) and poplar (269).

When the *R* paralogous genes were mapped along pear pseudomolecules, they were found to be non-randomly distributed across all 17 chromosomes (**Supplementary Fig. 18**). More than 30% of *R* genes were clustered in groups (**Supplementary Fig. 19)**, and clusters were most abundant on chromosomes 2, 5, and 11 (**Supplementary Fig. 20)**. Enrichment of *R* genes in

these corresponding genomic regions indicated that resistance gene evolution might involve tandem duplication and divergence of linked gene families, similar to those found in other known plant genomes.

_S_-locus comparisons in GSI self-incompatible species

Pear, exhibits typical gametophytic self-incompatibility (GSI) controlled by an apparently single multi-allelic locus (the *S*-locus; de Nettancourt 1997) containing at least two linked genes, one is a pistil *S*-determinant, known as *S-RNase* gene (Ushijima et al. 2003), and the other is a pollen *S*-determinant, proposed as an *S*-haplotype-specific F-box protein (SFB) and identified in *Prunus* species (Wu et al. 2009; Zhang et al. 2007). However, candidates for *SFB* genes controlling pollen self-incompatibility in pear have remained unclear until now. Based on the assembled sequence of the pear genome, the *S*-locus is anchored close to the end region of 3.7 M-4.6 M of LG17, which is consistent with its location along the genetic map of pear (Yamamoto et al 2007). Altogether six candidate *SFB* genes within the *S*-locus are predicted and show high frequencies of amino acid polymorphisms, ranging from 61.7% to 76.1%. As the scaffold containing *S-RNase* genes in pear is unanchored, an accurate estimation of the physical distance between *S-RNase* genes and the six candidate *SFB* genes cannot be made (**Supplementary Table 11**). Comparisons of the *S*-locus region (1000 kb) for pear, apple, strawberry, and potato (*Solanum tuberosum* L.), show that there are relatively moderate levels of synteny in this region (**Fig. 5**), and few common genes, except for *S-RNase* and *SFB* genes, are present in different plant species. Gene trafficking and rearrangements are active in this region. These findings indicate that the evolution of the *S*-locus region must have occurred following divergence of Rosaceae.

Unlike apple and strawberry, the six candidate *SFB* genes are present as tandem repeats in

pear. Thus, we propose that this unique characteristic of pear may be the result of gene duplication, thereby suggesting that a different mechanism for pollen self-incompatibility may be involved. Another interesting finding is the detection of highly repetitive sequences in *S*-locus regions of pear, apple, and potato, but not in strawberry, which exhibits self-compatibility (**Fig. 5**). Suppression of recombination at the *S*-locus region may be related to presence of many repetitive sequences in pear. The function of repetitive sequences in GSI requires further studies. Moreover, different repetitive sequences may play a role in the evolution of the *S*-locus as reported in Brassica species (Fujimoto et al. 2006).

Biological processes underlying fruit quality

Stone cells - Stone or grit cells, present in flesh of fruit, are important features of fruit quality in pear, but they are rare in other fruits. Lignin is the primary component of stone cells in pear fruit (Tao et al. 2009), and its synthesis has direct influence on formation and content of stone cells, ultimately influencing quality of pear fruit. By annotating the lignin biosynthesis pathway, it is revealed that lignin metabolism related genes in pear have similar levels of abundance to those found in apple and poplar, where lignin is involved in wood formation (**Supplementary Table 12**). Following phylogenetic analysis, 66 lignin synthesis-related gene families in pear show expansion, with pear exhibiting a greater demand for lignin synthesis. Predicted functions of TFs involved in the lignin pathway have identified that gene numbers of NAC and LIM families (**Supplementary Table 8**), reported to be related to lignin synthesis (Zhong et al. 2006; Kawaoka et al. 2000), were more than those found in strawberry, grape, and papaya. These two forms of TFs may be involved in lignin formation in pear fruit.

To further pursue genomic analysis of lignin in pear fruit, RNAseq data from three stages of fruit development (S422, early development; S627, middle development; and S830, near

ripening) have been analyzed (**Fig. 6**). Genes involved in lignin synthesis are highly expressed in the first two stages, almost 10-fold higher than those detected at or near ripening. Expression levels of genes encoding hydroxycinnamoyl transferases (HCT) are high at early stages of fruit development. *HCT* genes are known to promote lignin synthesis (Hoffmann et al. 2004), as companions to highly expressed genes encoding *p*-coumaroyl-shikimate/quinate 3'-hydroxylases (C3'H) and caffeoyl-CoA O-methyltransferase (CCOMT), leading to high levels of conversion of *p*-coumaroyl-CoA (PCC) into caffeoyl-CoA (CFC) and feruloyl-CoA (FC), and resulting in accumulation of both G-lignin and S-lignin. These findings support the hypothesis that there are higher levels of G-lignin and S-lignin in pear stone cells, but not of P-lignin. Meanwhile, none of the caffeic acid 3-O-methyltransferase (COMT) genes are expressed at all three stages of fruit development, suggesting that the rate limiting step for synthesis of lignin in pear fruit is that of conversion of CFC into FC (**Fig. 6**).

Sugar - The composition and content of sugar has an important influence on fruit quality and flavor. Instead of sucrose, which is common in non-rosacaeous species, sorbitol is a major photosynthetic product and phloem-translocated component in rosaceous fruit crops. A comparison of sorbitol metabolism-related genes in different species has revealed that the three gene families of sorbitol transport (SOT), sorbitol dehydrogenase (SDH), and sorbitol-6-phosphate dehydrogenase (S6PDH) in pear are higher than those in non-rosaceous species, but similar to those found in apple and strawberry (**Supplementary Table 13, Supplementary Fig. 21**). This indicates that duplication of the whole sorbitol metabolism pathway may have occurred to promote species fitness (van Hoek and Hogeweg 2009).

Gene families of *S6PDH, SDH*, and *SOT* have been expanded in both pear and apple genomes, and all three gene families belong to the *Maloideae*-specific clade (**Supplementary**

Table 13). Despite the close relationship between pear and apple, notable differences can still be found in the number of *S6PDH* genes, four members in pear compared to 11 members in apple (**Fig. 7**). In addition, the four *S6PDH* genes in pear are clustered into two clusters on chromosomes 5 and 2; however, in apple, there is only a single cluster located on chromosome 10, with others scattered on different chromosomes and scaffolds. These findings demonstrate that *S6PDH* gene-expansion in apple or *S6PDH* gene-contraction in pear must have occurred following their divergence from a common ancestor. Moreover, transcriptome data indicate that all four *S6PDH* genes are expressed in fruit, thus indicating that sorbitol could also be re-synthesized from monosaccharides, especially during later stages of fruit development. A total of 15 *SDH* genes in pear are transcribed and clustered onto two homologous chromosomes, 1 and 7, along the same orientation. These are cross-paired on the phylogenetic tree (**Supplementary Fig. 22)**, indicating that *SDH* genes must have expanded mainly through whole genome duplication. Whereas, 15 *SDH* genes in apple are more scattered and oriented in different directions (**Supplementary Fig. 23)**, suggesting that potential transposition events must have occurred. In addition, presence of pear-specific and apple-specific *SOT* genes in the phylogenetic tree (**Supplementary Fig. 24)** suggests that *SOT* genes have continued to expand following their divergence from the common Rosaceae ancestor.

Volatiles – Aroma is another important trait of pear fruit quality. Volatile compounds are mainly derived from the metabolism of fatty acids, amino acids, and carbohydrates (Schwab et al. 2008). When comparing all genes involved in three likely pathways of different plant genomes, we have found that lipoxygenase (LOX) and alcohol dehydrogenase (ADH), both involved in the α-linolenic acid metabolism pathway, have higher numbers of genes in both pear

and apple (**Supplementary Table 14**). Further RNAseq data (**Supplementary Fig. 25)** have provided evidence that a third of *LOX* homologous genes are highly expressed during fruit development, reaching peak levels at the intermediate stage. Meanwhile, expression levels of *ADH* increased along with alcohol formation during fruit development. Therefore the metabolism of α-linolenic acid is likely to be important for aroma formation in pear.

The release of volatiles is another important aspect for perception of smell and flavor. The numbers of β-glucosidase homologous genes, that catalyze the release of aroma volatiles from glucose indicans, are high in both pear (101) and apple (158) (**Supplementary Table 14**). As not all 101 β-glucosidase genes have clustered with known β-glucosidase, there may be novel functional genes affecting aroma that are yet to be determined. Moreover, RNAseq data (**Supplementary Fig. 25)** have revealed that only 20% of β-glucosidase homologous genes are expressed in pear fruit, and their expression levels have declined during fruit development. This has indicated that low aroma, perceived by sensory evaluation in pear, may be attributed to presence of more volatiles in bound status that are not released.

CONCLUSIONS

The sequenced pear genome will expedite basic research and crop improvement of this fruit crop. Advances in next-generation sequencing technologies have allowed genome sequencing become accessible for crop plants; however, most perennial plant genomes are heterozygous, and assembling a heterozygous genome using WGS sequences is challenging and often results in inaccurate genome assembly. To overcome this limitation, a BAC-by-BAC approach is used in combination with the high throughput sequencing technology to limit cost while assuring the quality of the assembled genome. The BAC-by-BAC approach is labor-intensive, and for this project, a total of 76,608 Illumina sequencing libraries have been constructed, two libraries for

each of the 38,304 BACs. The high quality of this genome is demonstrated by accurate annotation of genes and correction of 16,041 misannotated genes and redundant scaffolds in the apple draft genome using the WGS strategy. As for the nine ancestral chromosomes reported in apple, we infer that these nine chromosomes are not only the origin of the Pyreae tribe, but also serve as the ancestors of the whole Rosaceae family.

The sequence of the pear genome provides an invaluable new resource for biological research of *Pyrus*. In this study, the pear genome and related transcriptome analysis have provided insights into mechanisms underlying important biological processes including stone cell formation, sugar accumulation, aroma formation and release. Availability of nearly all pear gene sequences should benefit researchers working on fruit quality, developmental controls and disease resistance by enabling genome-wide functional studies and accelerating identification of gene-trait associations. In addition, further genome-wide comparative studies will provide insight and advance our knowledge on the genome evolution of Rosaceae. The high collinearity between pear and apple, combined with strawberry, provides more opportunity to reveal significant microsynteny, and the availability of the genome sequence will enable continued comparative genomics studies among species that will shed new light on gene family evolution.

METHODS

Genome sequencing

A BAC-to-BAC strategy combined with WGS sequencing was employed in assembly of the genome sequence of pear. And we used Illumina Hiseq 2000 to sequencing the genome.

For BAC libraries construction, *Hin*dIII and *BamH*I were used to generate partially digested insert DNA and these were ligated into appropriate sites of the vector pSMART (Lucigen, USA).

Ligations were transformed into phage resistant *Escherichia coli* EPI-300 host cells. On average, insert sizes in these BAC libraries ranged between 80 Kb to 180 Kb. For BAC clone DNA isolation, following culturing single colony in LB medium with antibiotics and growing 16 to 20 hours at 37℃, DNA was extracted and digested with *Not* I, Then, pulsed-field gel electrophoresis was used to separate transformed DNA from *E coli*. The quality and quantity of BAC DNA were checked using UV-VIS spectrometer along with gel electrophoresis runs of random samples. Usually, at least 0.75 µg BAC DNA was necessary for a single library preparation.

For BAC sequencing library construction, an Agilent Bravo Automated Liquid Handling Platform (Agilent, USA) and an Agilent BenchCel Microplate Handler (Agilent, USA) were used. Initially, the Adaptive Focused Acoustics (AFA) DNA fragmentation system (Covaris, USA) was used to fractionate DNA samples based on insert sizes. For the automated batch processing capability, 96-microTUBE plates (Covaris) were used as sample vessels. Then, T4 DNA polymerase (Illumina) and *E. coli* DNA polymerase I Klenow fragment (Illumina) were used to convert overhangs resulting from fragmentation into blunt ends. To ligate index adapters (**Supplementary Table 16**), having single 'T' base overhangs at 3' ends of DNA fragments, the polymerase activity of Klenow fragment was used to add 'A' bases to 3' ends of blunt DNA fragments. Following ligation, DNA samples with different index adapters were pooled together, according to the sample's position on the plate. Then, unligated index adapters were removed along an electrophoresis gel, and DNA segments of particular sizes were selected. Subsequently, index primers (**Supplementary Table 16**) were ligated to DNA segments, and PCR was used to selectively enrich those DNA fragments having index adapters and index primers on both ends, and also to amplify the amount of DNA in the library. Then, gel electrophoresis was

used to remove unligated index primers and to select DNA segments based on size. Finally, quality control tests were conducted using Agilent 2100 Bioanalyzer (Agilent) and StepOnePlus Real-Time PCR System (ABI, USA). Prior to sequencing, 23 96-well plates were pooled into a single lane (i.e., 2208 samples/lane), resulting in an average throughput of ~100.8 M reads/lane (assuming a read length of 100 bp, adding up to ~10G/lane). After sequencing, contamination from *E. coli* reads (about 6%) were filtered out in raw data of each BAC prior to assembly.

For WGS sequencing library construction, a Illumina genomic DNA library construction protocol was used, and a total of 10 paired-end or mate-pair libraries, spanning sizes of 180 bp to 40 kb, were constructed **(Supplementary Table 2)**. Most reads generated from mate-pair libraries (insert size >=2000 bp) were in the order of 49 bp; whereas, the corresponding length of paired-end libraries (ranging from 180 to 800 bp) was 100 bp.

For transcriptome sequencing, fruit samples at 15 days, 80 days, and 145 days after flowering (DAF) were used. RNA sequencing libraries were constructed using an Illumina standard mRNA-Seq Prep Kit (TruSeq RNA and DNA Sample Preparation Kits v2).

Genome assembly and SNP calling

The pear genome was assembled using short-read assembly software *SOAPdenovo* (http://soap.genomics.org.cn/) and sequence alignment software BLAT (Kent 2002). Scaffolds were constructed using *SSPACE* (Boetzer et al. 2011) software. First, each BAC was assembled with K=27 by *SOAPdenovo* using pair-end reads (250 bp and 500 bp), and then WGS mate-pair reads were used to construct scaffolds by SSPACE. Later, assembled BAC sequences were mixed, and a seed sequence, ~3 kb at ends of each scaffold, was selected to perform BLAT (Kent 2002) alignment for all scaffold sequences. Subsequently, similar sequences were combined and filtered for redundant bases using alignment results. If ends of scaffolds were aligned at

high identity (90%), they were merged into a single scaffold. Whereas, if a short scaffold aligned at high identity to the interior of another scaffold, then the shorter scaffold was deleted. If sequences had mutual complementation, these were combined into a single scaffold. After several iterations of these steps, the whole sequence length tended to stabilize. Finally, scaffolds were further linked into super-scaffolds by mate-pair WGS reads (2 Kb to 40 Kb) using *SSPACE* software (Boetzer et al. 2011), and gaps were filled with short read data.

The quality of the assembly was assessed by alignment to Sanger-derived phase 5 BAC sequences. Using *nucmer* software (http:// mummer.sf.net/) to identify the scaffold related to the BAC, then BACs were aligned to scaffolds using BLAST (Altschul et al. 1990). The *SOAPaligner* (http://soap.genomics.org.cn/soapaligner.html) was used to map reads to BACs, and statistics were performed for each BAC.

WGS reads (insert size <2K) were aligned to the genome using Bwa (Li and Durbin 2009), and SoapSNP (http://soap.genomics.org.cn/soapsnp.html) was used to detect SNPs. Further filtering conditions were set as quality scores of the consensus genotype of more than 20, sequencing depth of the site of more than 4 and less than 120, and with an average copy number of a nearby region of less than 2.

Development of RAD markers and anchoring of scaffolds

Individual genetic maps, derived from an F1 population of a cross between 'Bayuehong' and 'Dangshansuli' and consisting of 102 individuals, were used to develop an integrated map for anchoring scaffolds.

Genomic DNA was isolated from fresh leaves using the plant genomics DNA Kit (TIANGEN, Beijing, China), according to the manufacturer's recommendations. The Restriction-site Associated DNA (RAD) protocol (Chutimanitsakun et al. 2011) was used, except for the use of

*EcoR*I (recognition site: 5'G^AATTC3'). A total of 24 F1 individuals were pooled into a sequencing library with nucleotide multiplex identifiers (4 bp, 6 bp, and 8 bp; About 1 Gb), and 50 bp-reads (9.94 Mb reads data for each progeny on average) were generated on the NGS Illumina platform HiSeq2000. SNP calling algorithm was done using a Stacks package (Catchen et al. 2011) with default parameters. SNP markers were filtered by testing against expected segregation ratios (1:2:1 or 1:1) using a *chi*-square test, and then their sequence reads were aligned to scaffolds by BLAT (Kent 2002). Only those unique aligned SNPs with a cutoff value of 90% identity were kept. Finally, all qualified SNP markers were used to construct the pear consensus map using CP population option and the Kosambi mapping function in JoinMap version 3.0 (Van Ooijen and Voorrips 2001).

Repeat sequences

The Repbase (Repbase16.02) (Jurka et al. 2005) was used to find repeats by using *RepeatProteinMask* (Smit et al. 2004) and *RepeatMasker* (Smit et al. 2004). *RepeatModeler* (Smit et al. 2004) was used to build *de novo* repeats. Then redundancies were filtered out, and *Repeatasker* (Smit et al. 2004) was used to identify positions of repeats. Through structural features, *LTR_FINDER* software (Xu and Wang 2007) and *TRF* software (Benson 1999) were used to find Long Terminal Repeats (LTRs) and tandem repeats, respectively.

For structure-based search of intact TEs, LTR retrotransposons were detected by LTR_Finder (Xu and Wang 2007) and LTR_STRUC (McCarthy and McDonald 2003). Insertion time of intact LTRs was estimated by computing with Dismat after measuring distances between 5' and 3' solo-LTRs. MITEs were detected by MITE_Hunter (Han and Wessler 2010). LINEs were detected by MGEScan_nonLTR (Rho and Tang 2009). DDE domain TEs were detected by

checking flanking sequence alignments of DDE domains which were identified by scanning plant TE domains.

Genome and ncRNA annotation

AUGUSTUS (Stanke et al. 2006) and *GlimmerHMM* (Majoros et al. 2004) were used to perform *de novo* prediction based on the repeat-masked genome. Homologous proteins of other plant species (apple, strawberry, grape, and *Arabidopsis*) were mapped to the genome using *TblastN* (Altschul et al. 1990) with an E-cutoff value of *1e*-5. Aligned sequences, as well as their corresponding query proteins, were then filtered and passed to *GeneWise* (Birney et al. 2004) to accurately search for spliced alignments. Then, *GLEAN* (Mackey et al. 2005) was used to integrate these two sources of evidence to produce a consensus gene set.

About 7.8 Gb transcriptome sequence, of mixed multiple tissues generated by Illumina RNA-seq, was used to predict transcripts with Tophat (http://tophat.cbcb.umd.edu) and Cufflinks (http://cufflinks.cbcb.umd.edu). Then, predicted transcripts were used to complement the GLEAN gene set or were integrated as isoforms. Novel genes were added to generate the final gene set.

tRNAscan-SE (Lowe and Eddy 1997) was performed to search for reliable tRNA positions. Searches for snRNA and miRNA were done through a two-step method - first aligned with *Blast* and then searched with *INFERNAL* against Rfam database (Griffiths-Jones, et al. 2005). rRNAs were detected by aligning with BlastN against known plant rRNA sequences.

Gene families and phylogenetic analyses

Proteins of pear, apple, strawberry, grape, papaya, poplar, rice, and *Arabidopsis* were selected to perform all-against-all comparison using BLASTP (Altschul et al. 1990). The results were fed into the stand-alone OrthoMCL (Li et al. 2003) program using a default MCL inflation

parameter of 1.5. Single-copy families were selected to perform alignment by MUSCLE (Edgar 2004). Four-fold degenerate sites (4d) were picked by PhyML (Guindon et al. 2010), based on the maximum likelihood method (Guindon and Gascuel 2003), to reconstruct the phylogenetic tree using rice as an out-group. The divergence time was estimated by MultiDivtime (Edgar 2004) using the divergence time of papaya and *Arabidopsis,* identified by fossil records (Crepet et al. 2004). Subsequently, CAFÉ (De Bie et al. 2006) was used to identify gene family expansion and contraction.

Collinearity and WGD

MCscan (Tang et al. 2008) was used to identify collinearity blocks using paralog gene pairs, which were then identified by BLASTP (Altschul et al. 1990). Through MUSCLE alignment of gene pairs in collinearity blocks (Edgar 2004), 4dTv (transversion of four-fold degenerate site) values of each block were calculated using the sum of transversion of four-fold degenerate sites divided by the sum of four-fold degenerate sites (Huang et al. 2009). Ks values were calculated using MUSCLE (Edgar 2004) alignment and PAML (Yang 2007) for gene families of paralogous gene pairs. All these values were used for WGD analysis.

Disease resistance genes

Identification of pear resistance-related genes was based on the most conserved motif structures of plant resistance proteins, including CC, KIN, TIR, NBS, and LRR finger domains. Conserved motifs were derived from domain profiles retrieved from PFAM, PANTHER, PRINTS, PROSITE, SMART, and SUPERFAMILY databases, and from PAIRCOIL2 (McDonnell et al. 2006) to specifically detect CC domains. 'Other' types of *R* genes (without most conserved motifs, nevertheless potential *R* genes) were determined by BLAST (Altschul et al. 1990) based on a threshold of 60% similarity using the 'Other'-type reference of PRGDB (Sanseverino et al. 2010)

as a reference sequence. Assigning candidate genes to different *R* classes was based on the aforementioned protein domain composition. *R* genes were grouped into clusters when they were not interrupted by more than eight other ORFs (open reading frames) encoding non-R proteins.

ACCESSION CODES

This Whole Genome sequences of the pear (*Pyrus bretschneideri* Rehd.) project has been deposited at DDBJ/EMBL/GenBank under the accession AJSU00000000. The version described in this paper is the first version Pbr_v1.0, AJSU01000000. The data also posts at our professional website (http:// peargenome.njau.edu.cn).

AUTHOR CONTRIBUTIONS

S-L. Zhang, J. Wu and J. Wang managed the project.
S-L. Zhang, J. Wu, Z-B. Shi, S. Zhang, Z-W. Wang, R. Ming, M. A. Khan, S. S. Korban, X. Xu and T. Nishio designed the analyses.
J. Wu, K-J. Qi, H. Yin, C. Gu, Y-N. Yang, Y-T. Wang, W-J. Huang, M-S. Dai, Y-Z. Wang and S-J. Zhang collected samples and prepared DNA and RNA.
Z-W. Wang, J. Wu, S-L. Zhu, S-T.Tao, X. Zhao, L. Cong, J. L. Bennetzen, N. Chen, R. E. Paull, H. Wang, X-W. Wang, C. Deng, Y-H. Sha, D-H. Shi, H. Yin, J-Y. Wu, G-H. Qin, H-P. Zhang, S-F. Sha, C-L. Wang, Y-Z. Wang, C-Y. Gou, W-L. Zhou, D-L. Zhan, J. Wang, M-S. Chen, G-B. Wang, L. Xie, J-Y. Wang and S-L. Zhang contributed to sequencing, sequence assembly, genome annotation, genome structure, evolution and pathway analyses.
J. Wu, H. Chen, Y. Song, Y. Tao, L. Zeng, L-T. Li, D. Zheng, M. A. Khan, M-Y. Zhang, J-M. Sun, L-L. Xu, Y. Li, X. Liu, Q-S. Li, J-H. Shen and S-L. Zhang contributed to genetic mapping and chromosome anchoring.
J. Wu, Z-W. Wang, S-L. Zhu, X-S. Huang, Valpuri Sovero, R. E. Paull, S. S. Korban, M. A. Khan, R. Ming and S-L. Zhang wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (30900974, 31071759, 31000888, 31171928 and 31171936), the Fundamental Research Funds for the Central Universities (KYZ200911, KYZ201146), and the earmarked fund for China Agriculture Research System (CARS-29).

References

Al-Dous E. K., George B., Al-Mahmoud M.E., Al-Jaber M. Y., Wang H., Salameh Y. M., Al-Azwani E. K., Chaluvadi S.,et al.2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotech*.29:521-527

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Bennetzen, J. L., Schmutz J., Wang H., Percifield R., Hawkins J., Pontaroli A. C., Estep M., Feng L., Vaughn J. N., Grimwood J., Jenkins J., Barry K., Lindquist E., Hellsten U., et al. 2012. Full genome sequence analysis of the model plant Setaria. *Nature Biotech*. 30:555-561.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Birney, E. 2010. Assemblies: the good, the bad, the ugly. *Nat Methods* **8:** 59-60.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and genomewise. *Genome Res* **14:** 988-995.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27:** 578-579.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1:** 171-182.

Chutimanitsakun, Y., Nipper, R.W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E.A., and Hayes, P.M. 2011. Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics* **12:** 4.

Crepet, W.L., Nixon, K.C., and Gandolfo, M.A. 2004. Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits. *Am J Bot* **91:** 1666-1682.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22:** 1269-1271.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. 2004. Open source clustering software. *Bioinformatics* **20:** 1453-1454.

de Nettancourt, D. 1997. Incompatibility in angiosperms. *Sex Plant Reprod* **10:** 185-199.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Fawcett, J.A., Maere, S., and Van de Peer, Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc Natl Acad Sci USA* **106:** 5737-5742.

Fujimoto R., Okazaki K., Fukai E., Kusaba M., Nishio T. 2006. Comparison of the genome structure of the self-incompatibility (S) locus in interspecific pairs of *S* haplotypes. *Genetics* 173:1157-1167

Griffiths-Jones S., Moxon S., Marshall M., Khanna A., Eddy S.R., Bateman A. 2005 Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33, D121-124.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696-704.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307-321.

Han, Y. and Wessler, S.R. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38:** e199-e199.

Hoffmann, L., Besseau, S., Geoffroy, P., Ritzenthaler, C., Meyer, D., Lapierre, C., Pollet, B., and Legrand, M. 2004. Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *Plant Cell* **16:** 1446-1465.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., and Ni, P. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* **41:** 1275-1281.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., and Jubin, C. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449:** 463-467.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K., and Ebinuma, H. 2000. Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. *Plant J* **22:** 289-301.

Kent, W.J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Khurana, P. and Gaikwad, K. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793-800.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**:1754-60.

Li, L., Stoeckert Jr, C.J., and Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178-2189.

Lombard P.B. and Westwood, M. N. 1987. Pear rootstocks. In: Rom RC, CarlsonRF(eds) Rootstocks for fruit crops. JohnWiley and Sons, New York, USA, 145–183

Lowe T.M. and Eddy S. R.1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 0955-964.

Mackey, A., Liu, Q., Pereira, F., and Roos, D. 2005. GLEAN: Improved eukaryotic gene prediction by statistical consensus of gene evidence. *Genome Informatics*.

Majoros, W., Pertea, M., and Salzberg, S. 2004. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20:** 2878-2879.

McCarthy, E.M. and McDonald, J.F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19:** 362-367.

McDonnell, A.V., Jiang, T., Keating, A.E., and Berger, B. 2006. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22:** 356-358.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., and Lewis, K.L.T. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452:** 991-996.

Rho, M. and Tang, H. 2009. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res* **37:** e143-e143.

Rubtsov, G. 1944. Geographical distribution of the genus *Pyrus* and trends and factors in its evolution. *Amer Natur* **78:** 358-366.

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L., and Ercolano, M.R. 2010. PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res* **38:** D814-D821.

Schwab, W., Davidovich-Rikanati, R., and Lewinsohn, E. 2008. Biosynthesis of plant-derived flavor compounds. *Plant J* **54:** 712-732.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., and Mane, S.P. 2010. The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* **43:** 109-116.

Smit, A.F.A., Hubley, R., and Green, P. 2004. RepeatMasker Open-3.0.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* **34:** W435-W439.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28:** 2731-2739.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. 2008. Synteny and collinearity in plant genomes. *Science* **320:** 486-488.

Tao, S., Khanizadeh, S., Zhang, H., and Zhang, S. 2009. Anatomy, ultrastructure and lignin distribution of stone cells in two *Pyrus* species. *Plant Sci* **176:** 413-419.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., and Salamov, A. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596-1604.

Ushijima, K., Sassa, H., Dandekar, A.M., Gradziel, T.M., Tao, R., and Hirano, H. 2003. Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *Plant Cell* **15:** 771-781.

van Hoek, M.J.A. and Hogeweg, P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol* **26:** 2441-2453.

Van Ooijen, J. and Voorrips, R. 2001. JoinMap® 3.0, Software for the calculation of genetic linkage maps. *Plant Research International, Wageningen, The Netherlands***:** 1-51.

Varshney R.K., Chen W., Li Y., Bharti A. K ., Saxena R. K., Schlueter J. A. 2012. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotech*.**30**:83-89

Vavilov, N.I. 1951. The origin, variation, immunity and breeding of cultivated plants. *Soil Sci* **72:** 482.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., and Pruss, D. et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat Genet* **42:** 833-839.

Wu, J., Gu, C., Zhang, S., Zhang, S., Wu, H., and Heng, W. 2009. Identification of S-haplotype-specific S-RNase and SFB alleles in native Chinese apricot (*Prunus armeniaca* L.). *J Hort Sci Biotech* **84:** 645-652.

Xu, Z. and Wang, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35:** W265-W268.

Yamamoto T., Kimura T.,Terakami S., Nishitani C., Sawamura Y., Saito T., Kotobuki K., Hayashi T. 2007.Integrated reference genetic linkage maps of pear based on SSR and AFLP markers. *Breed. Sci.* 57:321-329

Yang, L. and Bennetzen, J. L. 2009. Distribution, diversity, evolution and survival of *Helitrons* in the maize genome. *Proc. Natl. Acad. Sci. USA* **106**:19922-19927.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.

Zeven, A.C. and Zhukovsky, P.M. 1975. Dictionary of cultivated plants and their centres of diversity, centre for agricultural publishing and documentation, Wageningen, The Netherlands 62-63.

Zhang, S.L., Huang, S.X., Kitashiba, H., and Nishio, T. 2007. Identification of S-haplotype-specific F-box gene in Japanese plum (*Prunus salicina* Lindl.). *Sex Plant Reprod* **20:** 1-8.

Zhong, R., Demura, T., and Ye, Z.H. 2006. SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. *Plant Cell* **18:** 3158-3170.

Table 1. Summary of genome assembly features and annotation of the pear (*Pyrus bretschneideri* Rehd.) genome sequence.

Unit of assembly	Proportion/ Unit type	Number	Size	% assembly	N50 (kb)	Longest (Mb)
Contigs	All	25312	501.3 Mb	97.9	35.7	0.3
Scaffolds	All	2103	512.0 Mb	100	540.8	4.1
	Anchored	796	386.7 Mb	75.5	698.0	4.1
Genes	Total	42812	118.8 Mb	23.2		
	Exon	202169	50.2 Mb	9.8		
	Intron	159357	61.2 Mb	11.9		
ncRNA	miRNA	297	37168 bp	0.01		
	tRNA	1148	86791 bp	0.02		
	rRNA	697	228388 bp	0.04		
	snRNA	395	45301 bp	0.01		
Repetitive sequences			271.9 Mb	53.1		

Figure legends

Figure 1. Distribution of basic genomic elements of pear. (A), chromosome karyotype; colored segments are in accordance with the Rosaceous ancestor. (B), Gene density; the rate of sites within gene region per 100 Kb ranges from minimum 0 to maximum 0.8 illustrated by red line. (C), DNA TE (transposon elements) density; the rate of sites within DNA TE region per 100 Kb ranges from 0 to 0.65 illustrated by blue line. (D), RT TE (retrotransposon elements) density; the rate of sites within RT TE regions ranges from 0 to 1 illustrated by purple. (E), SNP density; the rate of SNP per 100Kb ranges from 0 to 0.03 illustrated by green. (F), GC content; the rate of GC content ranges from 0.25 to 0.45 illustrated by black. Circos (http://circos.ca) was used for constructing this diagram.

Figure 2. Comparisons between apple and pear for repetitive elements.
The major repeats in apple and pear revealed that genome size differences of apple and pear were mainly attributed to repeat sequences.

Figure 3. Distribution of four-fold degenerate site (4dTv) distances of duplicate gene pairs in pear, apple, and strawberry. A total of 1,085 synteny blocks in pear (726 in apple and 262 in strawberry) are selected to calculate 4dTv values. The distribution of 4dTv values in pear (in blue) and those of apple (in red) are similar, while those of strawberry (in black) are different, with a single peak around 0.65, suggesting that there is no recent whole-genome duplication (WGD) in strawberry. The green groups are synteny blocks (557) between pear and apple, revealing that these groups are more close to the y-axis, and suggesting a more recent divergence event must have occurred between pear and apple.

Figure 4. The evolutionary scenario of nine chromosomes of the Rosaceae ancestor. Pear and apple have the same chromosome karyotypes and same chromosomal evolution mode. The Pyreae tribe went through a recent whole-genome duplication (WGD). The Amygdaleae tribe, such as strawberry in the Rosoideae subfamily, has no recent WGD, but has chromosome fragmentation and recombination from nine to seven. It is estimated that the ancestor of Rosaceae had nine chromosomes. To demonstrate the evolutionary process from the eudicot ancestor to the Rosaceae ancestor, strawberry was compared with grape.

Figure 5. Genes and repeat sequences surrounding candidate *S-RNase* genes in pear, apple, strawberry and potato. The dashed line with dot endpoints connects candidate *SFB* genes of pear and apple that share the highest sequence identity (which are shown in Supplementary Table 10). The grey arrow indicates genes supported by experimental evidence. The vertex of the triangles is the 3' orientation of the particular gene: red, candidate *S-RNase* gene; purple, candidate *SFB* gene; blue, other genes in the neighboring regions; green, repeat sequence.

Figure 6. (A) The phenylpropanoid biosynthesis pathway in pear that influences the conformation of stone cells in fruit. Red box indicates genes that had detectable expression, light red shaded ovals indicate important intermediate compounds in lignin pathway, green text and arrows show pathways with minor expression, and blue boxes show three important end-product compounds in pear fruit and green-boxed compound cannot be detected.

(B) Transcript ratio distribution of all enzymes related to lignin. Three stages of fruit development (S422, early development; S627, middle development; and S830, near ripening) were assessed. The ratio of S422 and S627 is shown along the x-axis and the ratio of S830 and S627 is shown along the y-axis. For points, different colors correspond to different enzymes, while different shapes correspond to different conditions of false discovery rate (FDR) values.
Abbreviations of genes involved in phenylpropanoid biosynthesis pathway are as follows: LP, L-Phenylalanine; CAN, Cinnamic acid; PCA, P-Coumaric acid; CA, Caffeic acid; FA, Ferulic acid; 5HA, 5-Hydroxyferulate acid; SA, Sinapic acid; CNC, Cinnamoyl-CoA; PCC, p-Coumaroyl-CoA; CFC, Caffeoyl-CoA; FC, Feruloyl-CoA; 5HC, 5-Hydroxyferuloyl-CoA; SC, Sinapoyl-CoA; PCouA, p-Coumar aldehyde; CafA, Caffeyl aldehyde; ConA, Conifer aldehyde; 5HydA, 5-Hydroxyconifer aldehyde; SinA, Sinapoyl aldehyde; PCAlc, p-Coumaryl alcohol; CFAlc, Caffeyl alcohol; CNAlc, Coniferyl alcohol; 5HydAlc, 5-Hydroxyconiferyl alcohol; SinAlc, Sinapyl alcohol; PHL, p-Hydroxyphenyl lignin; GL, Guaiacyl lignin; 5GL, 5-Hydroxy-guaiacyl lignin; SL, Syringyl lignin; PAL, phenylalanine ammonia-lyase; C4H, trans-cinnamate 4-monooxygenase; 4CL, 4-coumarate--CoA ligase; HCT, shikimate O-hydroxycinnamoyltransferase; C3'H, coumaroylquinate 3'-monooxygenase; COMT, caffeic acid 3-O-methyltransferase; CCOMT, caffeoyl-CoA O-methyltransferase; F5H, ferulate-5-hydroxylase; CCR, cinnamoyl-CoA reductase; CAD, cinnamyl-alcohol dehydrogenase; and POD, peroxidase.

Figure 7. Phylogenetic relationships, distribution patterns, and transcriptional expression of *S6PDH* genes. The phylogenetic tree was constructed using the Maximum Likelihood Method with Mega 5.0 software (Tamura et al. 2011). Heatmaps of expression patterns were drawn using Cluster 3.0 (de Hoon et al. 2004) along with expression levels (FPKM) of each of the *S6PDH* genes. Different colors have been used for different species. S422, S627, S830 are three different stages of development. Dotted lines between circles correspond to deleted non-*S6PDH* genes.

Fig.1

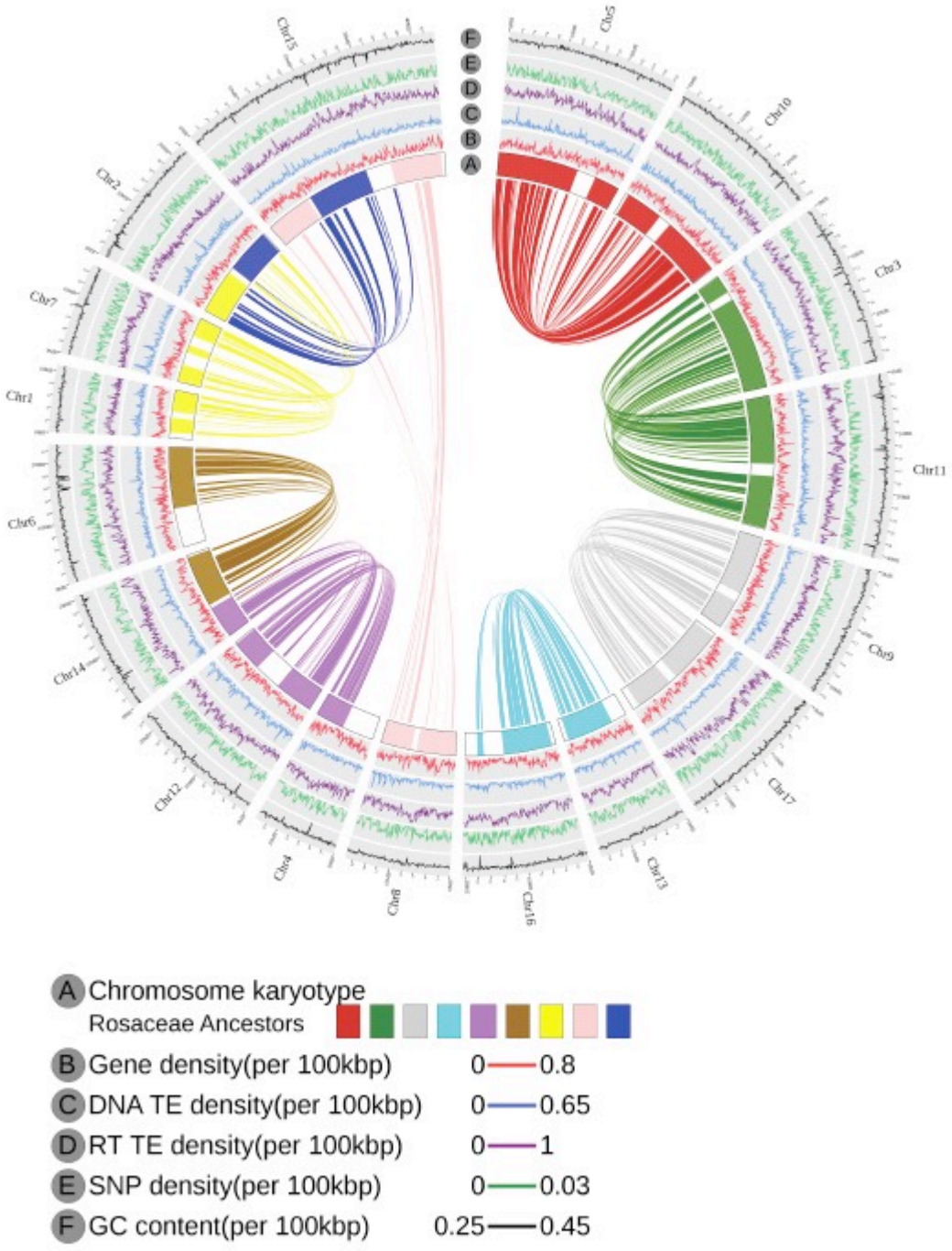


Fig.2

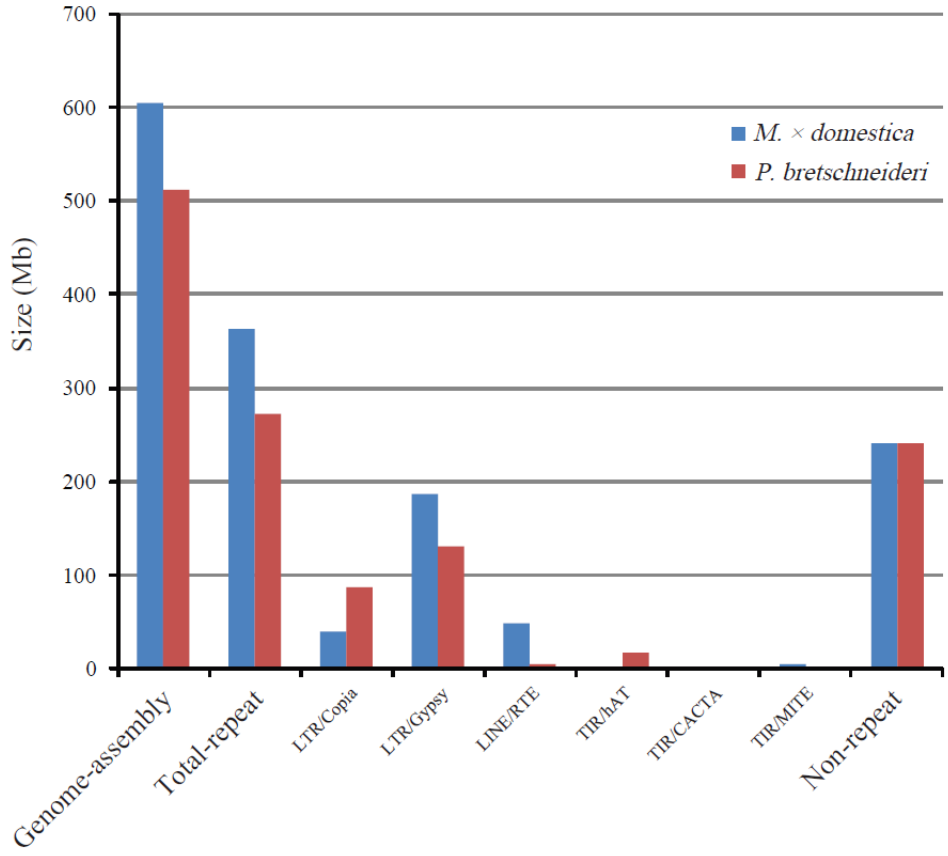


Fig.3

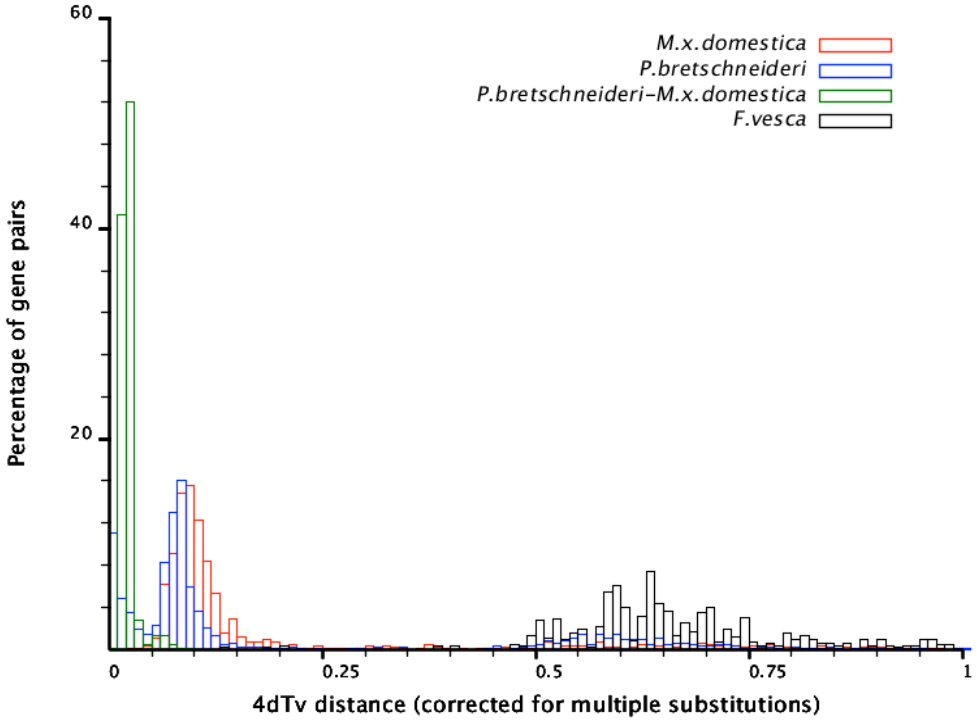


Fig.4

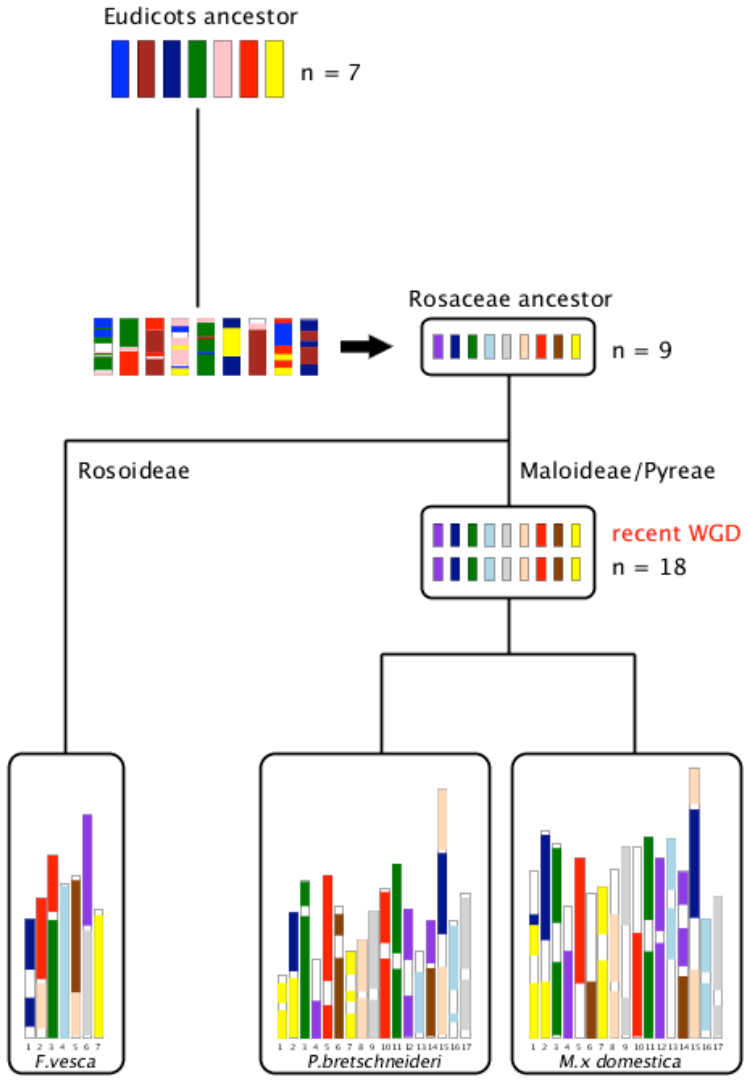


Fig.5

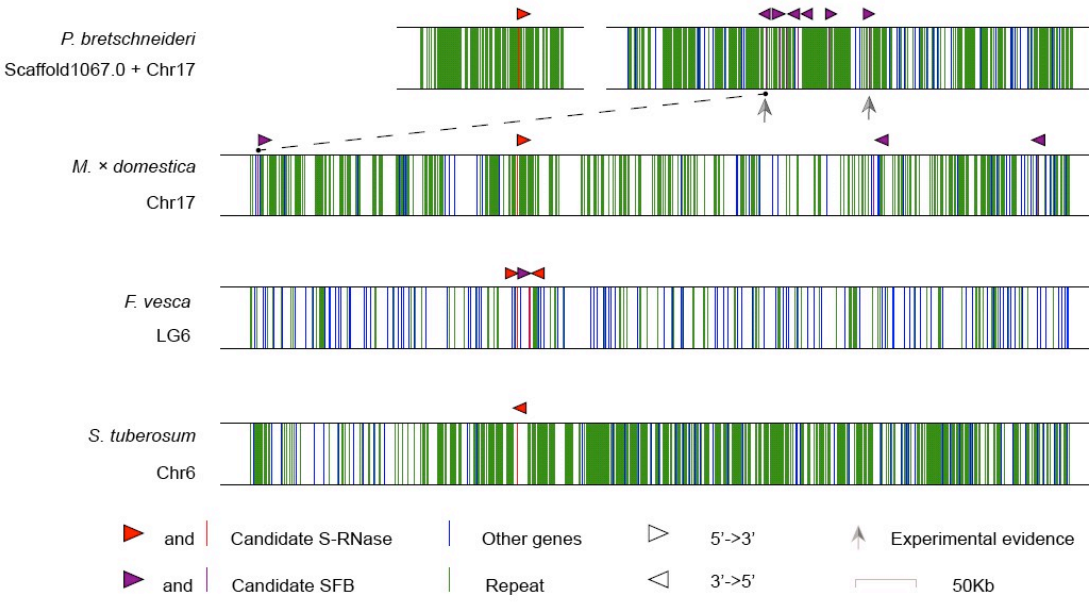
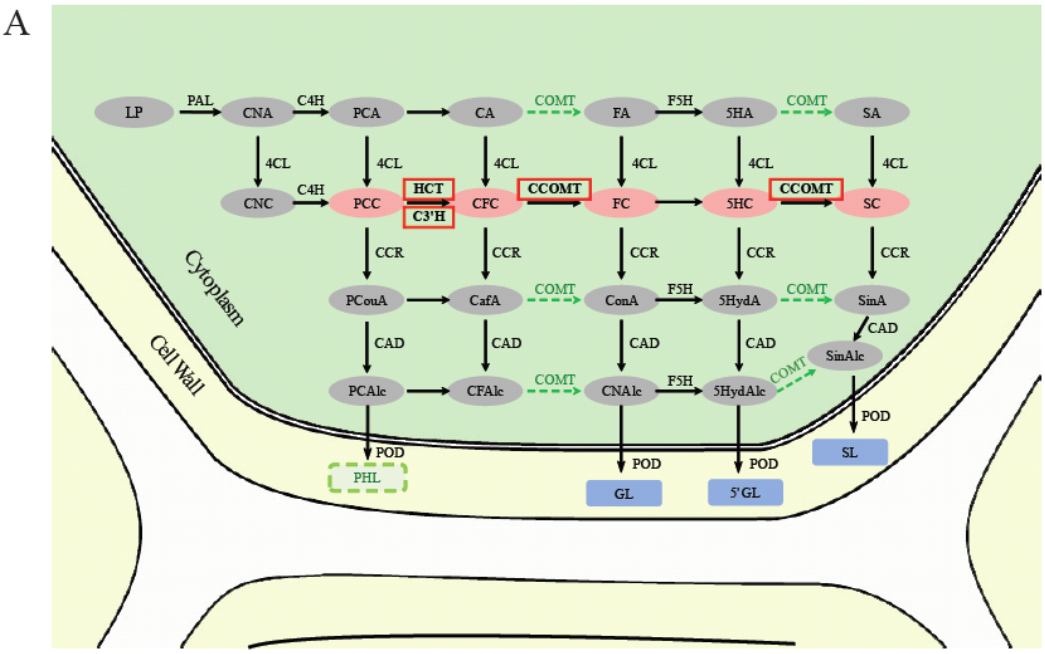


Fig.6



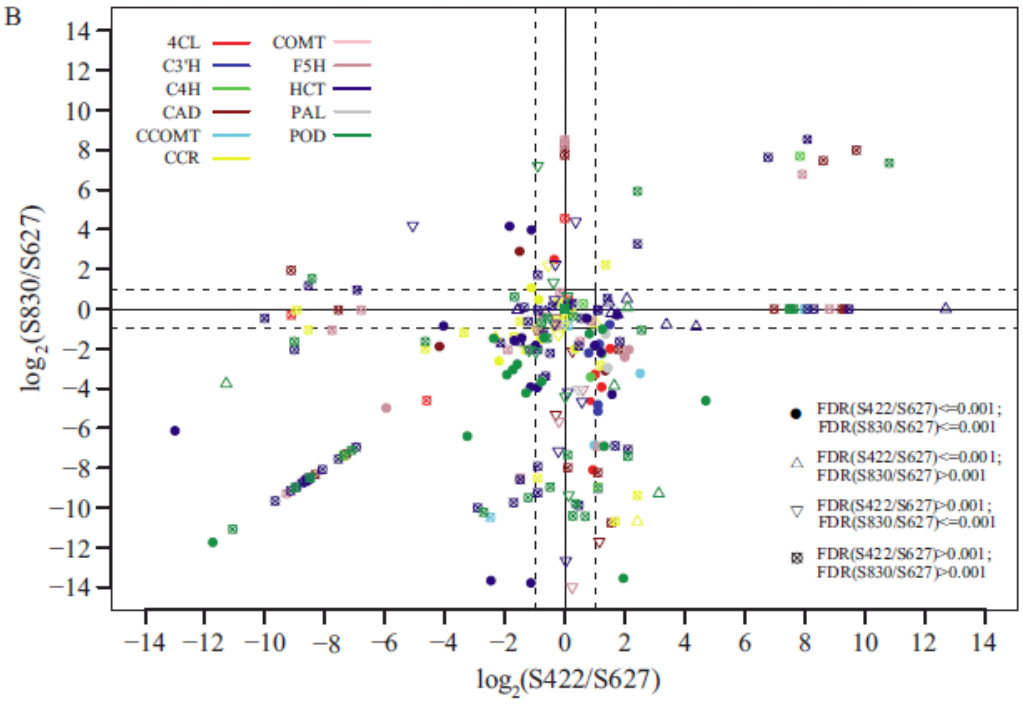


Fig.7