

Qixiang Zhang and Lidan Sun

Abstract

The genome of *Prunus mume* (*mei*), which was domesticated in China more than 3000 years ago as an important fruit and ornamental plant, was one of the first sequenced genomes among the *Prunus* sub-families of Rosaceae. In this study, the 280 M genome was assembled into scaffolds by combining 101-fold NGS data and optical mapping data; 83.9% of these scaffolds were further anchored to eight chromosomes in a genetic map constructed by restriction site-associated DNA (RAD) sequencing. Combining the *P. mume* genome data with other available genome data, we reconstructed nine ancestral chromosomes of the Rosaceae family, depicting chromosome fusion, fission and duplication history in the three major Rosaceae subfamilies. We sequenced the transcriptome of various tissues and performed a genome-wide analysis to reveal the

characteristics of *P. mume*, including its regulation of early blooming in endodormancy and biosynthesis of flower scent. The *P. mume* genome sequence adds to our understanding of Rosaceae evolution and provides important data for the improvement of fruit trees.

5.1 Plant Material

Wild samples of *P. mume* were sequenced by the Illumina Genome Analyzer II in Tongmai, Bomi County, Tibet, China, which is the western-end region of the origin of domesticated *P. mume* (Fig. 5.1).

The genetic maps that were used to develop the integrated map for anchoring the scaffolds were derived from F_1 populations, totalling 260 individuals from the cross between ‘Fenban’ and ‘Kouzi Yudie’ from Qingdao Meiyuan.

5.2 Methods

5.2.1 DNA Preparation and Whole-Genome Shotgun Sequencing

We used a whole-genome shotgun sequencing strategy with the Illumina Genome Analyzer. Total DNA was extracted from fresh young leaves of wild specimens from Tongmai Town, Tibet,

Q. Zhang (✉) · L. Sun

Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Engineering Research Center of Landscape Environment of Ministry of Education, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, School of Landscape Architecture, Beijing Forestry University, Beijing 100083, China
e-mail: zqx@bjfu.edu.cn

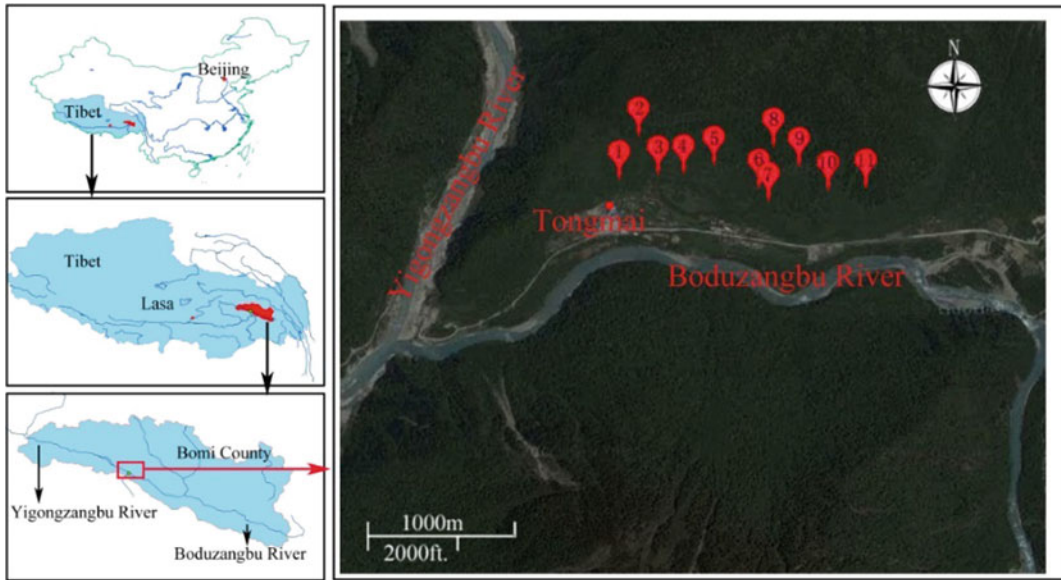


Fig. 5.1 Sample distribution map of GPS for *P. mume* in Tongmai town, Tibet. Number 4 represents the location of the sample used for *P. mume* sequencing

China, using the plant genomic DNA extraction kit (TIANGEN, Beijing, China) in accordance with the manufacturer's instructions. To obtain sufficient DNA to construct and sequence the library, we performed whole-genome amplification. We constructed eight paired-end sequencing libraries with insert sizes of approximately 180 base pairs (bp), 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb. In total, we generated 50.4-Gb data of paired ends, ranging from 50 to 100 bp. These data were checked and filtered on the reads that were generated previously (Li, et al. 2010). We filtered the low-quality reads using the following filtering criteria: reads had an 'N' over 10% of their length; reads contained more than 40-bp low-quality bases (quality score ≤ 5); reads contained more than 10-bp adapter sequences (allowing ≤ 2 -bp mismatches); small insert size paired-end reads that were overlapped (≥ 10 bp); Read1 and Read2 of paired-end reads were completely identical, which means that these paired-end reads were determined to be artefacts of the PCR experiment. Using stringent criteria, 28.4-Gb high-quality data were used for de novo genome assembly.

5.2.2 K-mer Analysis

We determined the relationship between sequencing depth and the copy number of a certain K-mer (refers to a sequence with K base pairs e.g. 17-mer), and when ignoring the sequence error rate, heterozygosity rate and repeat rate of the genome, the K-mer of the distribution should follow the Poisson theoretical distribution. The size of the genome was estimated using the total length of the sequence reads divided by the sequencing depth; the peak value of the frequency curve represents the overall sequencing depth. We estimated the genome size as $(N \times (L - K + 1) - B)/D = G$, where N is the total number of sequence reads, L is the average length of sequence reads and K is the K-mer length, defined as 17 bp. Here, B is the total number of low-frequency (frequency ≤ 1 in this analysis) K-mers, while G is the genome size and D is the overall depth, estimated from the K-mer distribution. It must be pointed out that as the K-mer of the distribution should approximate the Poisson distribution, not all low-frequency

K-mers will be errors. This might lead to an underestimate of the genome size, especially at low sequencing depths.

5.2.3 Genome Assembly

We performed a whole-genome assembly using SOAPdenovo (Version 1.05) (Li et al. 2010) with high-quality reads. Reads were loaded into the computer memory, and de Bruijn graph data structure was used to represent the overlap among the reads. The graph was simplified by removing erroneous connections and solving tiny repeats by read path. On the simplified graph, we broke the connections at repeat boundaries and output the unambiguous sequence fragments as contigs. Before generating scaffolds, in the ‘map’ step, SOAPdenovo realigned all usable reads to the contig sequences and obtained aligned paired ends (PEs). Subsequently, the software calculated the number of shared PE relationships between each pair of contigs, weighed the rate of consistent and conflicting PEs and constructed the scaffolds step by step, from short-insert to long-insert PEs. To close the gaps inside the constructed scaffolds, which were composed primarily of repeats that were masked before scaffold construction, we used the PE information to retrieve the read pairs that had one end mapped onto the unique contig and the other located in the gap region and performed local assembly using GapCloser (Version 1.12) with these collected reads.

5.2.4 Estimation of Heterozygosity Rate

The heterozygosity rate was calculated by calling the heterozygous SNPs. All high-quality reads were mapped onto the genome assembly using the software SOAP2 (<http://soap.genomics.org.cn/soapaligner.html>) with a cut-off of less than five mismatches. Subsequently, the alignment results were analysed for SNP mining using the SOAPsnp (<http://soap.genomics.org.cn/soapsnp.html>). The sites that met the following criteria

were searched and termed criterion-effective sites: (a) quality score of consensus genotype in the SNP mining result is greater than 20; (b) counts of all the mapped best and second best base are supported by at least four unique reads; (c) sequencing depth is more than 10X; (d) SNPs are at least 5 bp away from each other, with an additional requirement to the criterion-effective sites that the number of reads-supported best base is smaller than four times the number of reads-supported second best base (reads-supported best base/reads-supported second best base < 4) were identified as heterozygosity sites. Finally, the heterozygosity rate was estimated as the number of heterozygosity sites divided by the number of criterion-effective sites.

5.2.5 Whole-Genome Mapping

High-quality DNA (high molecular weight, >200 kb), specific for whole-genome mapping, was prepared from fresh *P. mume* leaves. Whole-genome shotgun single-molecule restriction maps were generated with the automated Argus system (OpGen Inc., Maryland, USA). First, DNA molecules were deposited onto silane-derivatised glass surfaces in MapCards (OpGen Inc., MD, USA) and digested with *Nhe* I or *Bam*H I for 20 min. Then, the DNA was stained with JOJO fluorescent dye (Invitrogen, CA, USA) and imaged with the Argus system. Overall, 243,174 single-molecule restriction maps (SMRMs) (>250 kb) with an average size of 344 kb were generated. Total size was approximately 83.6 Gb.

When using whole-genome mapping data to extend scaffolds, the original scaffold sequences were first digested in silico to generate corresponding restriction maps for each scaffold. Then, the in silico restriction maps were used as seeds to identify single molecules by length-based alignment using the Genome-Builder™ software package (<http://www.opgen.com/products-and-services/software/genome-builder>). These single-molecule maps were assembled together to extend the scaffolds with consensus restriction maps. Meanwhile, the scaffolds with

low coverage regions at both ends were trimmed, and only the high-quality extensions remained.

To extend sufficient scaffolds, this alignment-assembly process was iterated four times. All pairwise alignments were considered initial candidates for scaffold connection. The alignments with the highest scores remained when conflicts occurred. Super-scaffolds were constructed using the adjacent overlapping relationship between scaffolds; simultaneously, the orientation between each pair of scaffolds could be determined. The details on scaffold alignment and orientation can be obtained from the manufacturer of OpGen.

5.2.6 Identification of RAD Markers

The genetic maps that were used to develop the integrated map for anchoring the scaffolds were derived from F_1 populations, totalling 260 individuals from the cross between ‘Fenban’ and ‘Kouzi Yudie’ from Qingdao Meiyuan. Young leaves of these *P. mume* seedlings and their parents were collected for DNA extraction. Genomic DNA was isolated from the leaves using the Plant Genomics DNA kit (TIANGEN, Beijing, China) according to the manufacturer’s recommendations.

The RAD protocols were the same as in Chutimanitsakun et al. (2011), except we used *EcoR* I (recognition site: 5’G[^]AATTC3’). All 24 F_1 plants were pooled into one sequencing library with nucleotide multiplex identifiers (4, 6 and 8 bp). Approximately, 830 Mb of 50-bp reads (3.1 Mb of reads data for each progeny on average) were generated on the NGS platform HiSeq 2000. The SNP calling process was performed using the SOAP2 + SOAPsnp pipeline.

5.2.7 Genetic Map Construction and Scaffold Anchoring

A total of 260 F_1 seedlings of the cross between ‘Fenban’ and ‘Kouzi Yudie’ were used to construct the linkage map. Linkage analysis was performed using JoinMap version 3.0 (Van

Ooijen and Voorrips 2001). The RAD-based SNP markers were first tested against the expected segregation ratio. Two heterozygous SNP alleles between two parents were expected to segregate at a 1:2:1 ratio. One heterozygous and one homozygous SNP allele between two parents were expected to segregate at a 1:1 ratio. Distorted markers ($p < 0.01$) were filtered to construct a genetic map by the chi-square test. Subsequently, reads that contained SNP markers were aligned to the scaffolds. Only unique aligned SNPs with a cut-off of 87.5% identity remained per Blat (coverage ≥ 0.90) (Kent 2002). An LOD score of 12.0 was initially set as the linkage threshold for linkage group identification. Eight linkage groups that had the same number of *P. mume* chromosomes were formed at an LOD threshold of 12.0. All SNP markers were used to construct the *P. mume* consensus map with the CP population model in JoinMap, version 3.0.

To reduce the complex of scaffolds that were anchored to hundreds of SNP markers, a tag SNP was selected from each scaffold with multiple SNPs. We calculated the recombination fractions between all pairs of SNPs in a scaffold and chose the SNP that had the minimum recombination fraction in the sum. Tag SNPs were used to identify the order of scaffolds. Subsequently, two marginal SNPs were used to orient the scaffolds. Scaffolds with only one SNP marker could be anchored but not oriented due to a lack of markers. One hundred scaffolds (18% of 567 scaffolds) were labelled ‘uncertain orientation’.

5.2.8 Identification of Repetitive Elements

There are two main types of repeats in the genome, tandem repeats and interspersed repeats. We used Tandem Repeats Finder (Version 4.04) (Benson 1999) and Repbase (composed of numerous transposable elements, Versions 15.01) to identify interspersed repeats in the *P. mume* genome. We identified transposable elements in the genome at the DNA and protein levels. For the former, RepeatMasker (Version

3.2.7) was applied using a custom library (a combination of Repbase, a de novo transposable element library of the *P. mume* genome). For the latter, RepeatProteinMask, an updated tool in the RepeatMasker package, was used to conduct RM-BlastX searches against the transposable element protein database (Jurka et al. 2005). Identified repeats were classified into various categories.

5.2.9 Gene Prediction

To predict genes, four approaches were used: de novo prediction, the homology-based method, the EST-based method and transcript-to-genome sequences. For de novo prediction, Augustus (Stanke et al. 2006), GENSCAN (Salamov and Solovyev 2000) and GlimmerHMM (Majoros et al. 2004) were used with parameters trained on *Arabidopsis thaliana*. For the homology search, we mapped the protein sequences of four sequenced plants (*Cucumis sativus*, *Carica papaya*, *Fragaria vesca* and *A. thaliana*) onto the *P. mume* genome using TBLASTN, with an E-value cut-off of $1e^{-5}$; homologous genomic sequences were aligned against matching proteins using GeneWise (Birney et al. 2004) for accurate spliced alignments.

In the EST-based prediction, 4699 ESTs of *P. mume* were aligned against the *P. mume* genome using BLAT (identity ≥ 0.95 , coverage ≥ 0.90) to generate spliced alignments. The de novo set (28,610–36,095), four homology-based results (24,277–29,586) and the EST-based gene set (2001) were combined by GLEAN (Elsik et al. 2007) to integrate a consensus gene set. Short genes (CDS length < 150 bp) and low-quality genes (gaps in more than 10% of the coding region) were filtered. To finalise the gene set, we aligned RNA-Seq data from buds, fruits, leaves, roots and stems to the genome using TopHat (Version 1.2.0, implemented with bowtie1 Version 0.12.5) (Trapnell et al. 2009), and the alignments were used as input for Cufflinks (Trapnell et al. 2010) (Version 0.93) with default parameters. Open reading frames (ORFs) of those transcripts were

predicted using structure parameters trained on perfect genes from homology-based predictions. In the end, based on their coordinates on the genome sequences, we manually combined the GLEAN gene set and ORFs of transcripts to form the final gene set, which contained 31,390 genes.

5.2.10 RNA-seq Data Generation

Using TRIzol (Invitrogen), RNA was purified from five fresh tissues (bud, fruit, leaf, root and stem). The RNA sequencing libraries were constructed using the mRNA-Seq Prep kit (Illumina, San Diego, USA). Briefly, first-strand cDNA synthesis was performed with oligo-T primer and Superscript II reverse transcriptase (Invitrogen). The second strand was synthesised with *Escherichia coli* DNA Pol I (Invitrogen). Double-stranded cDNA was purified with a Qiaquick PCR purification kit (Qiagen) and sheared with a nebuliser (Invitrogen) to 100–500-bp fragments. After end repair and addition of a 3'-dA overhang, the cDNA was ligated to Illumina PE adapter oligo mix (Illumina) and size selected to 200 ± 20 -bp fragments by gel purification. After 15 cycles of PCR amplification, the 200-bp paired-end libraries were sequenced using the paired-end sequencing module (90 bp at each end) of the Illumina HiSeq 2000 platform.

5.2.11 Gene Annotation

Genes were aligned to the Swiss-Prot (release 2011.6) and TrEMBL (release 2011.6) databases using BLASTP ($1e^{-5}$) to determine the best match of the alignments. InterProScan (Version 4.5) motifs and domains of the genes were identified against protein databases of Pfam (release 24.0), PRINTS (release 40.0), PROSITE (release 20.52), ProDom (release 2006.1) and SMART (release 6.0). Gene ontology IDs for each gene were obtained by the corresponding InterPro entry. The genes were aligned against KEGG proteins (release 58), and the matches were used to establish the KEGG pathway.

5.2.12 Identification of Noncoding RNA Genes

The tRNA genes were predicted by tRNAscan-SE (Version 1.23) (Lowe and Eddy 1997). For rRNA identification, the rRNA template sequences (e.g. *A. thaliana* and rice) were aligned against the *P. mume* genome using BLASTN to identify possible rRNAs. Other noncoding RNAs, including miRNA and snRNA, were identified using INFERNAL (Version 0.81) by searching against the Rfam database (Release 9.1).

5.2.13 Comparative Genome Analysis

Paralogous and orthologous genes were identified by BLASTP search (E-value cut-off $1e^{-5}$). After removing self-matches, syntenic blocks (≥ 5 genes per block) were identified based on MCscan (Tang et al. 2008). The aligned results were used to generate dot plots; for self-aligned results, each block represents the paralogous region that arose from genome duplication, and for inter-aligned results, each block represents the orthologous region that was derived from a common ancestor. We calculated 4DTv (the number of transversions at fourfold degenerate sites) for each gene pair in the block and drew the distribution of 4DTv values to estimate the speciation between species or WGD events.

5.2.14 Identification of Duplicate and Syntenic Regions

Three new parameters were used to identify paralogous and orthologous relationships between *P. mume*, *Malus × domestica*, *F. vesca* and *Vitis vinifera* by BLASTN. Paralogous gene pairs that were identified during duplication analysis in *P. mume* and *M. × domestica*, respectively, and orthologous gene pairs that were identified by colinearity analysis between *P. mume* and *M. × domestica*, *F. vesca* and *M. × domestica*, *P. mume* and *V. vinifera* were validated by CloseUp (Hampson et al. 2004) analysis. Based on the

syntenic and duplication relationships, Rosaceae ancestral chromosomes were reconstructed, and the paleo-history was analysed.

5.2.15 Identification of CBF and BEAT Genes

The *CBF* genes of *P. mume* were identified with *A. thaliana CBF* genes using BLASTP (E-value $< 1e^{-10}$, identity $> 30\%$ and coverage $> 70\%$). The *BEAT* genes of *P. mume* were identified with *BEAT* genes (Gene Bank ID: AF043464) using BLASTP (E-value $< 1e^{-10}$, identity $> 30\%$ and coverage $> 70\%$).

5.3 Results

5.3.1 Sequencing and Assembly

To construct the reference genome of *P. mume*, we initially sequenced two domesticated *P. mume* samples using the Illumina Genome Analyzer II. Both samples had a high heterozygous rate, as estimated using K-mer statistics (Fig. 5.2a, b), and the de novo assembly results by current NGS algorithms did not meet the reference quality standards. We then considered wild samples due to their lack of artificial grafting and asexual reproduction. We collected wild samples of *P. mume* in Tongmai, Bomi County, Tibet, China, which is the western-end region of the origin area of domesticated *P. mume* (Xing et al. 2009). In a previous study (Chen 1995; Xing et al. 2009), we determined that the origin area of *P. mume* was confined to an area of approximately 0.7 km^2 , and that the primary distribution area was less than 0.3 km^2 , below an elevation of 2230 m. Samples from this region were highly homozygous due to generations of self-fertilisation in a hermetically sealed geographic environment. We chose one such wild sample for sequencing, and its low heterozygosity was confirmed by sequencing and K-mer statistics (Fig. 5.2c).

We generated 50.4 Gb of sequencing data for this wild sample, using the Illumina Genome

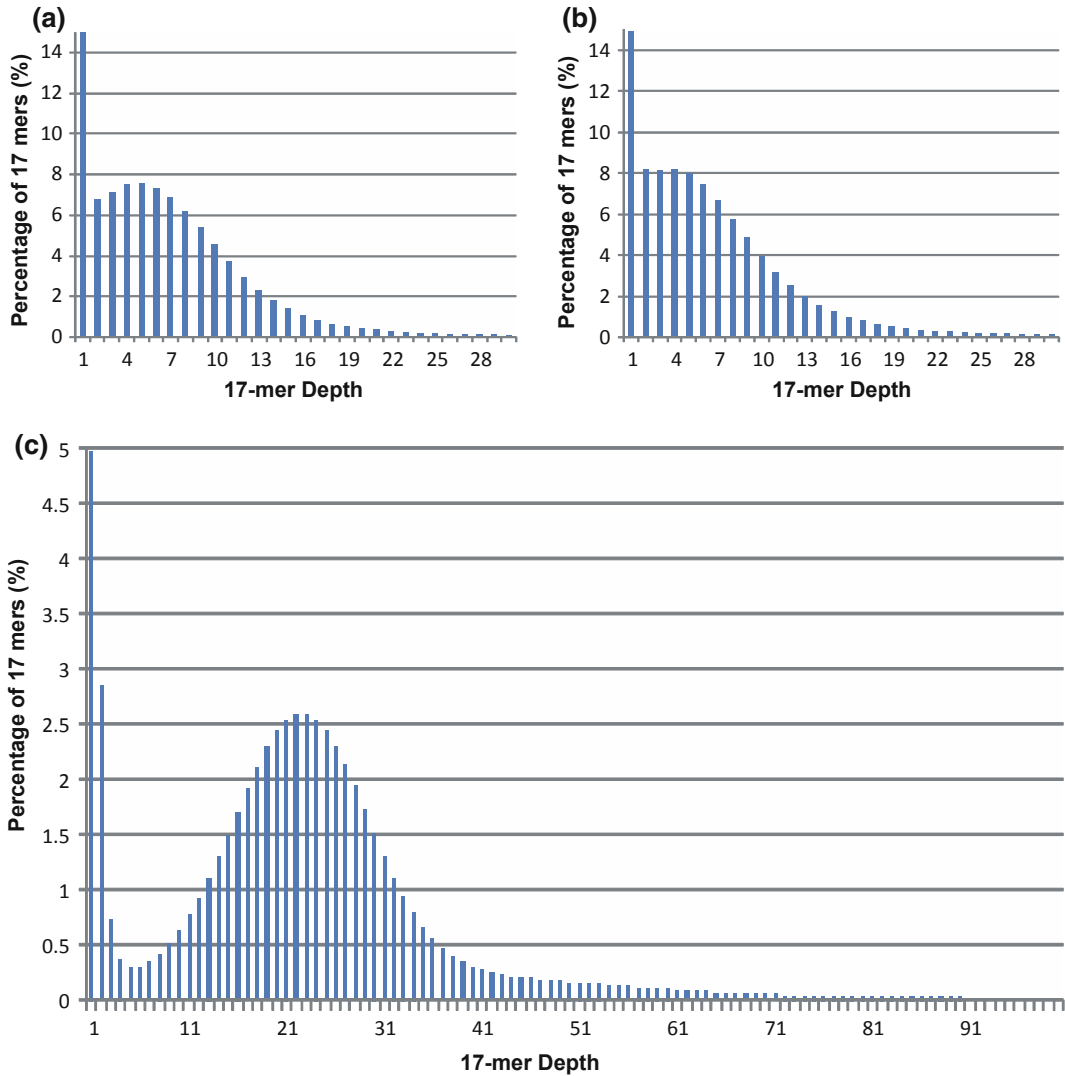


Fig. 5.2 K-mer analysis. **a, b** Estimating the domesticated samples, **c** Estimating the wild *P. mume* sample used for genome assembly. The x -axis represents depth (X); the y -axis is the proportion that represents the percentage at that depth. (Without consideration of the

sequence error rate, heterozygosity rate and repeat rate of the genome, the 17-mer distribution should obey the Poisson theoretical distribution. In the actual data, due to the sequence error, the low depth of 17-mer will take up a large proportion.)

Analyzer II, including three short-insert (180–800 bp) pair-end libraries and five large-insert (2–40 Kb) mate-pair libraries (Table 5.1). The SOAPdenovo (Li et al. 2010), a genome assembler algorithm that is based on the *de Bruijn* graph, was used to assemble the *P. mume* genome. Based on K-mer statistics, the *P. mume* genome was estimated to be 280 Mb (Fig. 5.2c). Approximately, 84.6% (237 Mb) of its genome

were assembled. The contig N50 of the assembled sequence was 31.8 Kb (longest, 201.1 Kb), and scaffold N50 was 577.8 Kb (longest, 2.87 Mb) (Table 5.2). By mapping raw reads back to the draft genome, we observed a heterozygosity rate of 0.03% in this wild sample, supporting our assumption of low heterozygosity in the wild sample. We identified 125,383-bp and 19,897-bp sequences in assembly similar to

Table 5.1 Construction of libraries, generation and filtering of sequencing data for genome assembly used

Library insert size (bp)	Read length (bp)	Raw data			Filtered data		
		Total data (Gb)	Sequence depth (X)	Physical depth (X)	Total data (Gb)	Sequence depth (X) ^a	Physical depth (X) ^a
180	100 PE	6.6	23.6	21.2	6.1	21.8	19.6
500	150 PE	10.2	36.4	60.8	7.1	25.4	42.2
800	100 PE	3.7	13.2	52.8	3.0	10.7	42.9
2000	45 PE	2.8	10.0	222.2	2.5	8.9	198.4
5000	45 PE	3.0	10.7	595.3	2.5	8.9	496.1
10,000	90 PE	11.4	40.7	2261.9	4.4	15.7	873.0
20,000	90 PE	4.7	16.8	1865.0	0.8	2.9	317.5
40,000	50 PE	8.0	28.6	11,442.9	2.0	7.1	2871.4
Total		50.4	180.0	16,522.1	28.4	101.4	4861.1

^aAssumed genome size is 280 Mb

Table 5.2 Statistics of *P. mume* genome assembly

	Contig		Scaffold		Whole-genome mapping	
	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number
N90	5769	7803	85,987	482	85,987	361
N80	12,180	5272	217,085	316	224,931	195
N70	18,473	3815	339,338	229	432,540	118
N60	24,813	2791	443,973	168	711,996	75
N50	31,772	2009	577,822	120	1085,026	48
Longest	201,075		2871,019		15,622,157	
Total number (>100 bp)		45,592		29,989		29,868
Total number (>2 Kb)		10,894		1449		1328
Total	219,917,886	45,811	237,149,662	29,989	237,166,662	29,868

chloroplast and mitochondrial sequences, respectively. Although most of the regions were shorter than 1000 bp, we found a ~25,000-kb chloroplast-similar region which might be the plasmid sequence or a plastid-transferred nuclear fragment.

To improve the assembly of the *P. mume* genome, we performed whole-genome mapping (WGM), an automated high-throughput optical mapping method (Zhou et al. 2004), to generate an entire genomic map (Fig. 5.3a). We constructed the WGM map with a 300-fold whole-genome depth using *Bam*H I and *Nhe* I independently. Through an iterative assembly

strategy, combined with WGM and sequence data, 170 scaffolds, that were assembled by NGS, were grouped into 49 large scaffolds; thus, the scaffold N50 improved significantly from 578 Kb to 1.09 Mb (Table 5.2).

Subsequently, we constructed a high-density genetic map by applying restriction site-associated DNA (RAD) marker strategy (Chutimanitsakun et al. 2011; Baxter et al. 2011; Miller et al. 2007) in a segregating F_1 population. The consensus genetic map contained eight linkage groups, consisting of a set of 1484 high-quality SNP markers (co-dominant markers), 779 of which were used in anchoring and

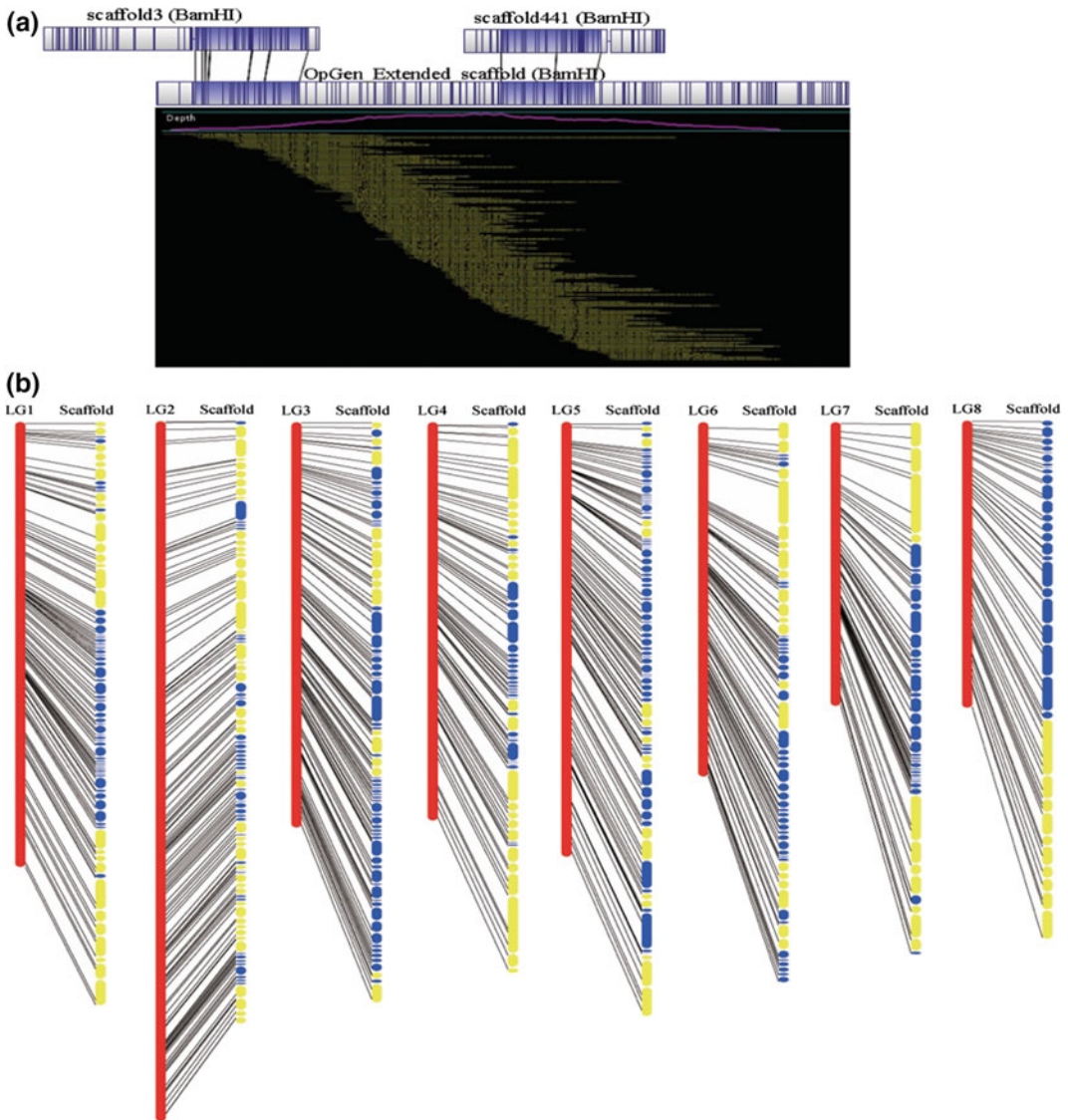


Fig. 5.3 Whole-genome mapping assembly and anchoring of the *P. mume* genome. **a** Assembly of *P. mume* genome by whole-genome mapping. **b** Anchoring of the *P. mume* genome into eight linkage groups using 779

high-quality SNP markers. Yellow scaffolds were anchored by whole-genome mapping and SNP markers, whereas the blue scaffolds were anchored by SNP markers

orienting scaffolds (Fig. 5.3b). The genetic map improves the quality of the reference and would be useful in map-based cloning and further marker-assisted molecular breeding.

5.3.2 Genome Annotation

We annotated 106.8 Mb (45.0% of the assembled genome) of repetitive sequences (Table 5.3)

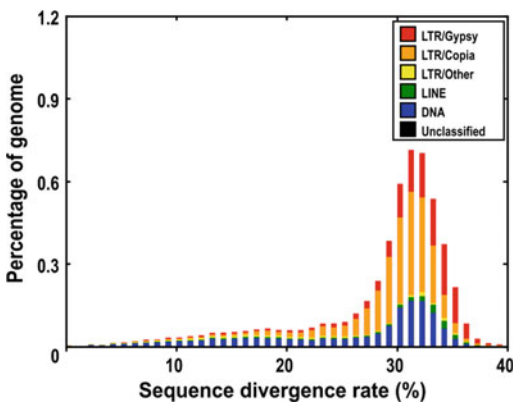
Table 5.3 Statistics of repeats in the *P. mume* genome

Type	Repeat size (Mb)	% of genome
Protein mask	17.32	7.29
RepeatMasker	12.36	5.20
Trf	10.58	4.45
Denovo	103.15	43.41
Total	106.75	44.92

in the *P. mume* genome by integrating the de novo and homology-based approaches. Transposable elements (TEs) were the predominant components, constituting 97.9% of all repetitive sequences. The long terminal repeat (LTR) *Copia* family and the *Gypsy* family are the most abundant TEs in the *P. mume* genome. The TE divergence rate suggested a lack of recent amplification (Fig. 5.4); these were conserved in *M. × domestica* and *F. vesca*, while *M. × domestica* had a much higher proportion in the *Gypsy* family than *P. mume* and *F. vesca* (Table 5.4).

To improve gene annotation, we generated 11.3 Gb of RNA-Seq data from five major tissues: bud, fruit, leaf, root and stem (Table 5.5). Using these data, integrated with ab initio homology prediction methods, we annotated 31,390 protein-coding genes, comparable with the value in *F. vesca* (34,809) (Shulaev et al.

2011) and less than that found for *M. × domestica* (57,386) (Velasco et al. 2010), as expected. Average transcript length in *P. mume* was 2514 bp, with 4.6 exons per gene (Table 5.6). Approximately 82.5% of all genes could be functionally annotated following a consensus method of either known homologous or predictive sequence signatures using Swiss-Prot, GO, TrEMBL (Bairoch and Apweiler 2000), InterPro (Zdobnov and Apweiler 2001) or KEGG (Kanehisa and Goto 2000), and we considered 98.3% of these annotations were high confidence (Table 5.7). We noted that 25,854 (82.6%) gene models were expressed, of which 768, 308, 240, 762 and 179 genes were expressed solely in buds, fruits, leaves, roots and stems, respectively. We further integrated 82,832 peach EST sequences; together with RNA-seq data, 85.1% of gene models represented transcripts. We also annotated the noncoding RNA genes in the current assembly, including 508 tRNA, 209 miRNA, 125 rRNA and 287 snRNA (Table 5.8).

**Fig. 5.4** Divergence rates of the transposable elements in the assembled scaffolds. The divergence rate was calculated based on the alignment between the RepeatMasker-annotated repeat copies and the consensus sequence in the repeat library

5.3.3 Genome Evolution

By genome self-alignment via MCscan (Tang et al. 2008) (Fig. 5.5) and 4DTv (the number of transversions at fourfold degenerate sites) distribution of duplicated pairs, there was no recent whole-genome duplication after *P. mume* species differentiation from *M. × domestica* (Fig. 5.6a).

We examined the paleo-history in *Prunus* and found that there was a triplicated arrangement (ancestral γ event). We aligned 27,819 gene models to the seven paleo-hexaploid ancestor chromosome groups in grape (Jaillon et al. 2007) and identified that the colinearity blocks

Table 5.4 Occurrence of transposable elements in sequenced Rosaceae genomes

Classification	<i>P. mume</i>			<i>M. × domestica</i>			<i>F. vesca</i>		
	Total length	TE coverage	Total genome coverage	Total length	TE coverage	Total genome coverage	Total length	TE coverage	Total genome coverage
	(Mb)	(%)	(%)	(Mb)	(%)	(%)	(Mb)	(%)	(%)
LTR/Copia	23.8	22.8	10	40.6	12.9	5.5	10.8	22.5	5.3
LTR/Gypsy	20.4	19.5	8.6	187.1	59.5	25.2	12.9	26.8	6.4
LTR/Other	21.8	20.8	9.2	3.2	1	0.4	8.5	17.7	4.2
LINE	3.1	3	1.3	48.1	15.3	6.5	0.7	1.5	0.3
SINE	0.9	0.9	0.4	-	-	-	0.2	0.4	0.1
DNA transposons	20.2	19.3	8.5	6.6	2.1	0.9	12.9	26.8	6.4
Other	1.1	1.1	0.5	-	-	-	2.1	4.4	1
Unknown	13.3	12.7	5.6	28.9	9.2	3.9	-	-	-
Total	104.6	100	44.1	314.5	100	42.4	48.1	100	

contained 2772 orthologs; the extent of these blocks covered 78.1% of the *P. mume* genome.

Further, we aligned 27,819 *P. mume* gene models to themselves and identified seven major blocks of duplication, which corresponded to 194 gene pairs, covering 38.5% of the anchored genome. The chromosome-to-chromosome relationships P5-P7, P2-P4-P8, P1-P2-P4-P6, P1-P5, P2-P5-P8, P2-P4-P7 and P3-P4 (Fig. 5.6b) suggested that triplicated arrangement (γ event) marks remained in the *P. mume* genome. Based on the evidence of paleo-hexaploidisation (γ event) and lineage-specific duplications in eudicots, it was possible to examine chromosomal changes during the evolution of *P. mume* and other Rosaceae species.

5.3.4 Reconstruction of Ancestral Chromosomes of Rosaceae

We reconstructed nine ancestral chromosomes of the Rosaceae family and determined the history of chromosome fusion, fission and duplication in the three major Rosaceae subfamilies. Previous studies have reported the eudicots ancestor with seven proto-chromosomes (Jaillon et al. 2007) and the grass ancestor with five proto-chromosomes

(Salse et al. 2009). Using *M. × domestica* as a reference, we analysed the syntenic relationships between the sequenced Rosaceae genomes of *P. mume* (P), *F. vesca* (F) and *M. × domestica* (M).

We identified 151 blocks that contained 4546 orthologous genes, covering 96.9% of the anchor *P. mume* genome, between *P. mume* and *M. × domestica* versus 132 blocks with 2031 orthologous genes, covering 88.8% of the *F. vesca* genome, between *F. vesca* and *M. × domestica*. The chromosome-to-chromosome orthologous relationships are shown in Fig. 5.7.

Combining intergenomic and intragenomic analysis of the Rosaceae genomes, we noted the following primary chromosome pair combinations: M5-M10/P3-P6/F2-F3, M3-M11/P1-P3-P6/F2-F3-F5, M9-M17/P2-P4/F1-F6, M13-M16/P2/F4, M4-M12-M14/P1-P2-P5-P8/F5-F6, M5-M14/P7/F5, M1-M2-M7/P5/F7, M8-M15/P2/F2-F5 and M1-M2-M15/P1-P2-P3-P8/F1 (Fig. 5.7). We reconstructed a putative ancestral genome of Rosaceae and proposed an evolutionary scenario of *P. mume*, *M. × domestica* and *F. vesca* from the putative nine-chromosome ancestor (Fig. 5.8).

In *P. mume*, chromosomes 4, 5 and 7 did not undergo rearrangement, coming directly from ancient chromosomes III, VII and VI,

Table 5.5 List of tissues and reads for whole-transcriptome sequencing mapped onto the *P. mume* genome

Tissue type	Bud	Fruit	Leaf	Root	Stem
Perfect match read no.	1,24,99,437	1,77,05,866	1,62,37,146	1,83,73,940	1,21,46,280
≤ 5 bp mismatch read no.	34,75,572	57,89,365	51,87,087	63,26,385	42,86,634
Unique match read no.	1,54,05,362	2,24,04,254	2,04,69,481	2,35,28,150	1,57,83,948
Multi-position match read no.	5,69,647	10,90,977	9,54,752	11,72,175	6,48,966
Total mapped reads no.	1,59,75,009	2,34,95,231	2,14,24,233	2,47,00,325	1,64,32,914
Total unmapped reads no.	35,23,231	53,38,431	45,09,397	66,38,567	38,28,296
Total reads no.	1,94,98,240	2,88,33,662	2,59,33,630	3,13,38,892	2,02,61,210
Total base pairs (bp)	1,75,48,41,600	2,59,50,29,580	2,33,40,26,700	2,82,05,00,280	1,82,35,08,900

respectively. Chromosome 1 came from ancestral chromosomes II, V and IX. Chromosome 2 originated from ancient chromosomes IV and VIII, into which some of chromosomes III and V were inserted. Chromosome 3 came from chromosomes I, II and IX; chromosome 6 came from I, II, and V, and chromosome 8 originated from ancient chromosomes V and IX. Thus, we hypothesised that at least eleven fissions and eleven fusions occurred in *P. mume* from the nine common ancestral chromosomes. For *M. × domestica*, at least one whole-genome duplication (WGD) and five fusions took place to reach the 17-chromosome structure, compared with 15 fusions for *F. vesca* to affect the 7-chromosome structure.

5.3.5 Early Blooming of *P. Mume*

P. mume is nearly the first tree that blooms in early spring, blooming even at temperatures below 0 °C. Thus, *P. mume* has a specific mechanism to acclimate to cold weather and to release itself from dormancy. The (DAM) dormancy-associated MADS-box transcription factors family, which is related to dormancy induction and release (Sasaki et al. 2011), was identified in the *P. mume* assembly, and all six *DAM* genes were noted in the arrayed tandem (Fig. 5.9a, Table 5.9). In a previous study, the authors identified six *DAM* genes, all of which were transcriptional repressors in *P. mume* (Sasaki et al. 2011). They also found that all *DAM* genes were repressed during prolonged exposure to cold and maintained at low levels until endodormancy release. The authors of another study observed that expression of *DAM3*, *DAM5* and *DAM6* was suppressed by chilly temperatures, bottoming on bud break in *Prunus persica* (Jiménez et al. 2010). These findings suggest that DAM inhibits dormancy release and that its expression is suppressed during prolonged exposure to cold to allow bud release from dormancy.

We hypothesised that the *DAM* genes explained the early dormancy release in *P. mume*. To this end, we examined the

Table 5.6 General statistics of predicted protein-coding genes

Gene set	EST	Protein homology search				Gene finder			GLEAN	RNA-Seq	Combine
		<i>C. sativus</i>	<i>C. papaya</i>	<i>F. vesca</i>	<i>A. thaliana</i>	Augustus	GENSCAN	GlimmerHMM			
Number	4,699	24,277	27,200	29,586	25,414	32,479	28,610	36,095	30,012	21,585	31,390
Average length of transcribed region (bp)	2001	2533	2022	2642	2412	2442	5211	2032	2523	2454	2514
Average length of CDS (bp)	562	1053	913	1043	1008	1175	1315	964	1164	1074	1146
#Exons per gene	3.1	4.2	3.7	4	4.2	5.1	6	3.9	4.7	4.4	4.6
Average length of exon (bp)	184	253	247	257	241	229	217	245	249	245	249
Average length of intron (bp)	701	469	411	521	441	307	772	364	369	409	380

Table 5.7 Functional annotation of predicted genes with homology or functional classification by each method

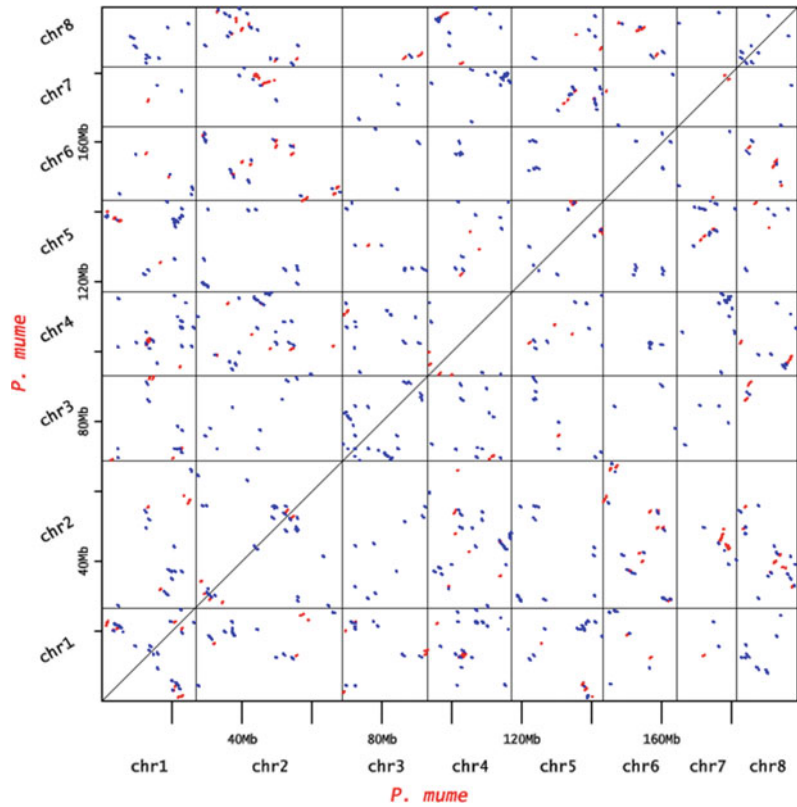
	Database	Number	Percent (%)
Annotated	Swiss-Prot	19,696	62.8
	InterPro	21,236	67.7
	GO	16,822	53.6
	KEGG	15,504	49.4
	Trembl	25,650	81.7
	Total	25,905 ^a	82.5
Unannotated		5485	17.5
Total		31,390	100

^a449 annotations were hits to hypothetical or uncharacterised proteins

Table 5.8 Noncoding RNA gene fragments in the current *P. mume* assembly

ncRNA Type	Copy#	Average length (bp)	Total length (bp)	% of genome
miRNA	209	120.65	25,216	0.0106
tRNA	508	75.21	38,209	0.0012
rRNA	125	196.89	24,611	0.0103
28S	46	348.98	16,053	0.0067
18S	17	111.29	1892	0.0008
5.8S	11	112.55	1238	0.0005
5S	51	106.43	5428	0.0022
snRNA	287	118.09	33,891	0.0142
CD-box	158	98.08	15,497	0.0065
HACA-box	21	118.14	2481	0.001
slicing	108	147.34	15,913	0.0067

Fig. 5.5 Whole-genome duplication in the *P. mume* genome mapped using gene collinear order information. Syntenic blocks are formed by red or blue dots, representing best hits across any two chromosomes in the same or opposite direction, respectively



phylogenetic relationships between *DAM* genes in *P. mume*, using the PHYML 3.0 software. The molecular evolution models of the six tandem *DAM* genes in *P. mume* suggested that they were derived from serial duplication events in the following order: *PmDAM1*, *PmDAM3*, *PmDAM2*, *PmDAM5*, *PmDAM4* and *PmDAM6* (Fig. 5.10). This model is consistent with previous studies in the peach genome (Jiménez et al. 2009), and we suggest that these duplication events are unique in the *Prunus* subfamily—we did not find these tandem *DAM* genes in *M. × domestica* or *F. vesca*. The application of two other programs, MEGA 4 (Tamura et al. 2007) and PAML4 (Yang 2007), obtained similar results. By estimation of pairwise dN and dS rates using MEGA 4, there was significant purifying selection and no significant positive

selection of the six sequences. In an analysis of *P. mume* *DAM* genes, most sites were highly conserved, with a dN/dS rate ratio near 0 or nearly neutral.

In addition to *DAM*, the *C-repeat binding transcription factor (CBF)* mediates the establishment of early dormancy release—overexpression of *PpCBF1* in *M. × domestica* results in a strong sensitivity to short day lengths and induced dormancy at optimal growth temperatures (Wisniewski et al. 2011). Horvath et al. (2010) noted that *EeDAM1* was cold stress-responsive and contains putative CBF-binding sites, which are cis-regulating motifs that are targeted by the cold/drought stress CBF regulon in the 2000-bp region upstream of the *EeDAM1* translation start codon. This finding suggests that *CBF* controls the cold-responsive *EeDAM1* gene (Horvath et al.

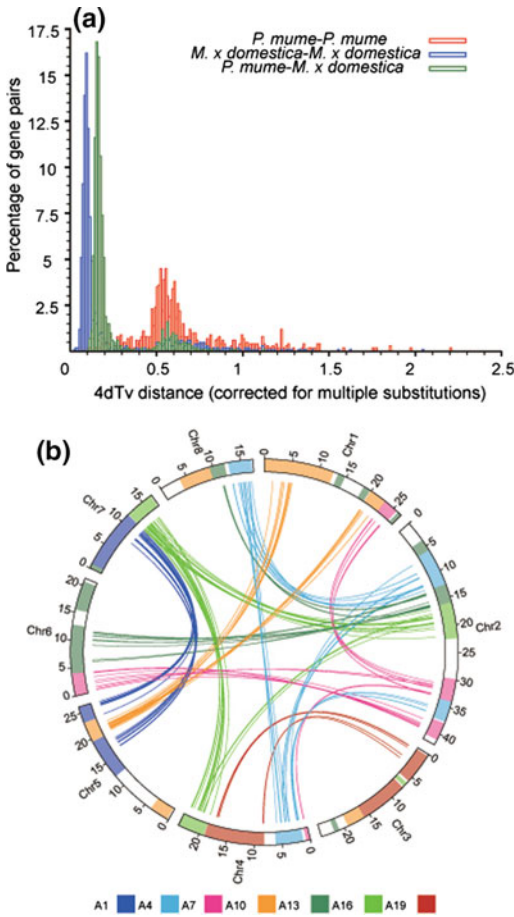


Fig. 5.6 Evolution of *P. mume*. **a** 4DTV distribution of duplicate gene pairs in *P. mume* and *M. x domestica*, calculated based on alignment of codons with the HKY substitution model. **b** The duplication of *P. mume* by paralogous pairs in the *P. mume* genome (chromosomes Chr1 to Chr8). Each line represents a duplicated gene. The seven colours reflect the seven ancestral eudicot linkage groups (A1, A4, A7, A10, A13, A16 and A19)

2010). Similar to *EeDAM1*, conserved *CBF* sites were identified in the 1000-bp region upstream of the translation start codons in *DAM4* to *DAM6* in peach and Japanese apricot (Sergio Jiménez et al. 2009; Sasaki et al. 2011). In the *P. mume* genome, we identified 13 *CBF* orthologous genes

(Tables 5.10) and seven *CBF* regulons, *late embryogenesis-abundant (LEA) proteins/dehydrins* (Table 5.11). In the upstream regions of *DAM* genes in *P. mume*, we noted more putative *CBF*-binding sites on *DAM4*, *DAM5* and *DAM6* than in peach and found novel sites, one on *DAM1* and two on *DAM6* (Fig. 5.9b). We suggest that these additional sites render *P. mume* more sensitive to cold and result in early blooming in spring. In summary, our analyses have increased our understanding of the molecular control of dormancy, flowering regulation and acclimation to cold. The *CBF* and *DAM* genes that we identified in *P. mume* might allow molecular biology facilities to breed fruits and ornamental plants with disparate blooming times.

5.3.6 Floral Scent

Floral scent, determined by a complex mixture of low-molecular-weight volatile molecules, has a significant function in the reproductive processes of many plants and enhances the aesthetic properties of ornamental plants (Pichersky and Dudareva 2007). In earlier research, we determined that the dominant compound classes in *P. mume* flowers were benzenoid/phenylpropanoid from the cinnamic acid pathway, in addition to terpene compounds.

Genes related to the biosynthesis of volatile compounds, such as (*BEAT*) benzyl alcohol acetyltransferase, which catalyses the synthesis of benzyl acetate (Dudareva et al. 1998; Aronovich et al. 2007), have been identified in the *P. mume* genome. The *BEAT* gene family expanded notably in *P. mume* (34 members) compared with *M. x domestica* (16), *F. vesca* (14), *V. vinifera* (4), *Populus trichocarpa* (17) and *A. thaliana* (3) (Table 5.12). Twenty-six of 34 *P. mume* *BEAT* genes lay in clusters, the largest of which contained 12 members that were

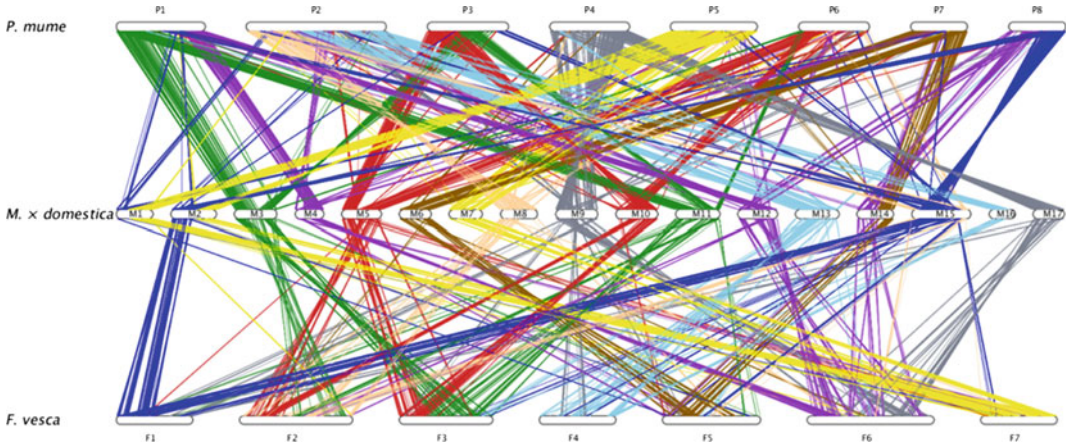


Fig. 5.7 Synteny between *P. mume*, *F. vesca* and *M. × domestica*. Schematic representation of the orthologs identified between *P. mume* (P1–P8), *F. vesca* (F1–F7) and *M. × domestica* (M1–M17). Each line represents

an orthologous gene. The nine different colours represent the blocks reflecting the origin from the nine ancestral Rosaceae linkage groups

arranged in tandem (Fig. 5.11), suggesting that *BEAT* genes originated from serial duplication events, in contrast to the other sequenced plants.

In summary, the expansion of the *BEAT* gene family might increase the content of benzyl acetate and be related to the special fragrance of *P. mume*. Research on the *P. mume* genome should allow us to breed novel aromatic cultivars and other aromatic plants in the Rosaceae family.

5.4 Discussion

We are eager to establish an appropriate reference genome for Rosid species, which include one-third of all flowering plants (Hummer and Janick, 2009). Currently, draft genome sequences are available for three model Rosaceae species—*M. × domestica*, *F. vesca* and *P. mume*. The domesticated apple (*M. × domestica*), the main fruit crop of temperate regions throughout the world, is highly heterozygous and has a large

genome, which creates technical challenges in assembling its genome, resulting in 1629 metacontigs.

The strawberry (*F. vesca*) has a much smaller genome of ~240 Mb, allowing functional gene studies within Rosaceae. Although the strawberry is useful for functional genomics research, most related high-value fruit plants in the Rosaceae species—peach (*P. persica*), pear (*Pyrus nivalis*) and cherry (*Prunus avium*)—are woody plants, not herbaceous ones. Unfortunately, the nearest relatives of woody fruit crops usually have a cumbersome polyploid genome.

After a detailed study of its origin and the current distribution in *P. mume*, we obtained a suitable sample from an isolated group in the origin area of *P. mume* for genomic sequencing and assembly. We report the genomic sequence of *P. mume* due to its small genome of ~280 Mb and low heterozygosity. With whole-genome mapping, we increased the scaffold quality to 1.1 Mb in N50 and constructed *P. mume* pseudochromosomes using 779

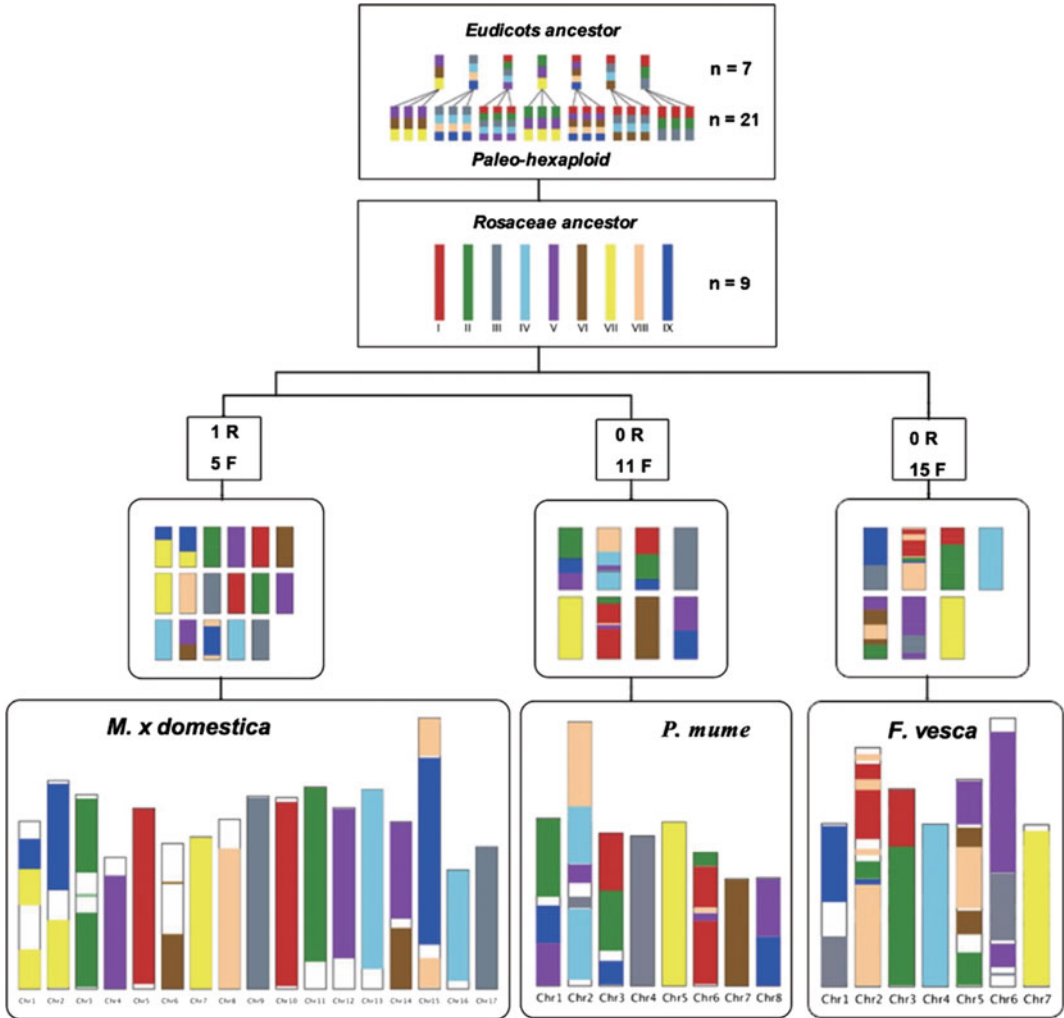


Fig. 5.8 Evolutionary model of the Rosaceae genome. The Rosaceae ancestor chromosomes are represented by nine colours. The various evolutionary processes from the common ancestor are indicated as R (whole-genome duplication (WGD)) and F (for fusions of chromosomes). In the second layer, different colours in each chromosome

represent the origin of the common ancestral chromosomes. The current structure of the Rosaceae genome is shown at the bottom of the figure. In some regions, we were not able to determine which ancestral chromosome they came from, and those regions therefore represented as white spaces

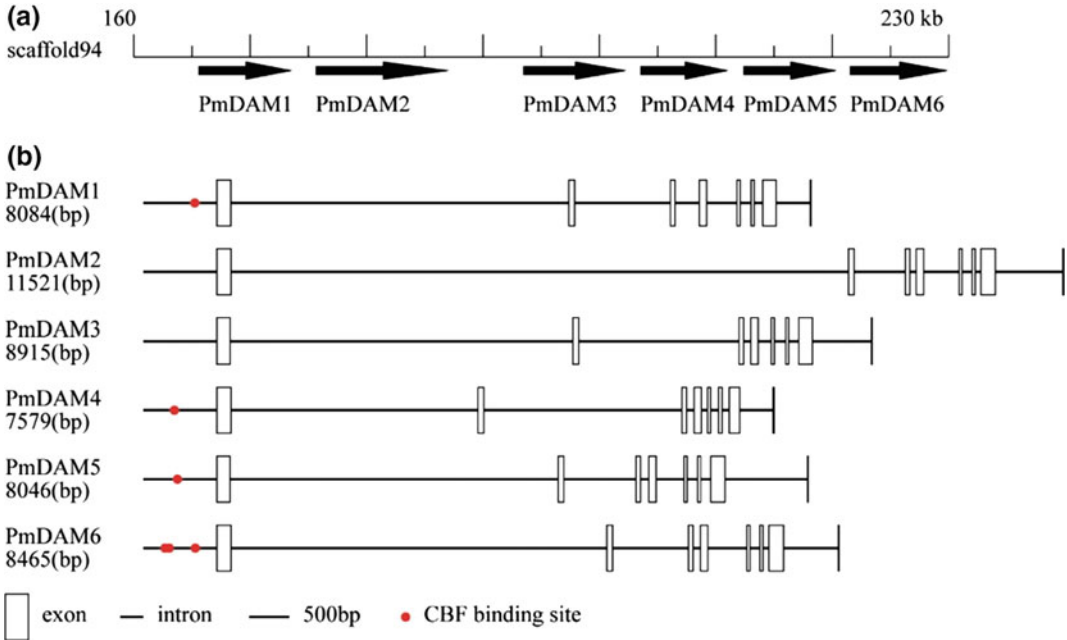


Fig. 5.9 Six tandemly arrayed *DAM* genes in *P. mume*. **a** Overview of *PmDAM* in the *P. mume* genome. Six *PmDAM* genes are located as tandem repeats. **b** Structures of *PmDAM* genes in *P. mume*. Boxes and lines represent exons and introns, whereas red points represent CBF-binding sites

Table 5.9 *DAM* gene orthologs of in *P. mume*

Gene name	Query species	ID	<i>P. mume</i> gene prediction	
			Scaffold	Genemark
<i>PmDAM1</i>	<i>Prunus persica</i>	gb DQ863253.2	scaffold94	Pm004420
<i>PmDAM2</i>	<i>Prunus persica</i>	gb DQ863255.1	scaffold94	Pm004419
<i>PmDAM3</i>	<i>Prunus persica</i>	gb DQ863256.1	scaffold94	Pm004418
<i>PmDAM4</i>	<i>Prunus persica</i>	gb DQ863250.1	scaffold94	Pm004417
<i>PmDAM5</i>	<i>Prunus persica</i>	gb DQ863251.1	scaffold94	Pm004416
<i>PmDAM6</i>	<i>Prunus persica</i>	gb AB437345.1	scaffold94	Pm004415

Fig. 5.10 Maximum likelihood rooted tree of six *P. mume* *DAM* genes; *PtMADS27* was used as outgroup

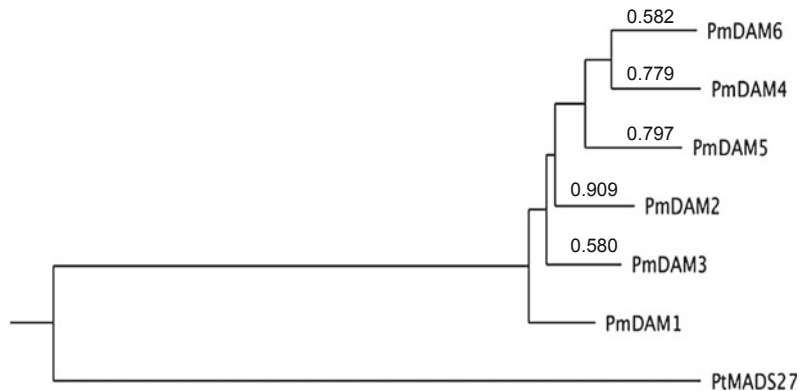


Table 5.10 CBF orthologs of in *P. mume*, *M. × domestica*, *F. vesca*, *P. trichocarpa*, *V. vinifera*, *O. sativa* and *A. thaliana*

Species	Number	Accession number
<i>P. mume</i>	13	Pm004870, Pm019385, Pm019386, Pm026227, Pm023766, Pm023767, Pm023768, Pm023769, Pm023770, Pm023772, Pm023775, Pm023777, Pm026221
<i>M. × domestica</i>	10	MDP0000154764, MDP0000155057, MDP0000189347, MDP0000195376, MDP0000198054, MDP0000262710, MDP0000400129, MDP0000451365, MDP0000652413, MDP0000833641
<i>F. vesca</i>	6	mrna13327.1, mrna13329.1, mrna30159.1, mrna30226.1, mrna32378.1, mrna32380.1
<i>P. trichocarpa</i>	14	POPTR_0001s08710.1, POPTR_0001s08720.1, POPTR_0001s08740.1, POPTR_0003s12120.1, POPTR_0004s19820.1, POPTR_0006s02180.1, POPTR_0009s14990.1, POPTR_0012s13870.1, POPTR_0012s13880.1, POPTR_0013s10330.1, POPTR_0015s13830.1, POPTR_0015s13840.1, POPTR_0016s02010.1, POPTR_0019s10420.1
<i>O. sativa</i>	11	Os01t0968800-00, Os02t0558700-00, Os02t0676800-01, Os02t0677300-01, Os03t0117900-01, Os04t0572400-00, Os08t0545500-00, Os09t0522000-01, Os09t0522100-00, Os09t0522200-02, Os11t0242300-00
<i>V. vinifera</i>	5	GSVIVT01019860001, GSVIVT01031387001, GSVIVT01031388001, GSVIVT01033793001, GSVIVT01033795001
<i>A. thaliana</i>	10	AT1G12610.1, AT1G12630.1, AT1G63030.1, AT2G35700.1, AT2G36450.1, AT4G25470.1, AT4G25480.1, AT4G25490.1, AT5G51990.1, AT5G52020.1

Table 5.11 Dehydrin orthologs in *P. mume*, *M. × domestica*, *F. vesca*, *P. trichocarpa*, *V. vinifera*, *O. sativa* and *A. thaliana*

Species	Number	Accession number
<i>P. mume</i>	7	Pm000687, Pm026682, Pm026683, Pm026684, Pm020945, Pm021811, Pm006114
<i>M. × domestica</i>	17	MDP0000126135, MDP0000129775, MDP0000178973, MDP0000196703, MDP0000265874, MDP0000269995, MDP0000360414, MDP0000529003, MDP0000595270, MDP0000595271, MDP0000629961, MDP0000689622, MDP0000698024, MDP0000770493, MDP0000862169, MDP0000868044, MDP0000868045
<i>F. vesca</i>	7	mrna14934.1, mrna14935.1, mrna14938.1, mrna14940.1, mrna17179.1, mrna21840.1, mrna27549.1
<i>P. trichocarpa</i>	8	POPTR_0002s01460.1, POPTR_0003s13850.1, POPTR_0004s16590.1, POPTR_0005s26930.1, POPTR_0009s12290.1, POPTR_0013s05870.1, POPTR_0013s05880.1, POPTR_0013s05890.1
<i>O. sativa</i>	7	Os01t0702500-01, Os02t0669100-01, Os11t0451700-00, Os11t0453900-01, Os11t0454000-01, Os11t0454200-01, Os11t0454300-01
<i>V. vinifera</i>	3	GSVIVT01018878001, GSVIVT01019440001, GSVIVT01023824001
<i>A. thaliana</i>	10	AT1G20440.1, AT1G20450.1, AT1G54410.1, AT1G76180.2, AT2G21490.1, AT3G50970.1, AT3G50980.1, AT4G38410.1, AT4G39130.1, AT5G66400.1

Table 5.12 Numbers of orthologous genes found in *P. mume* (Pm), *M. × domestica* (Md), *P. trichocarpa* (Pt), *A. thaliana* (At), *V. vinifera* (Vv), *F. vesca* (Fv) and *O. sativa* (Os) that synthesise volatile molecules

Type gene	Pm	Md	Pt	At	Vv	Fv	Os
<i>PAL</i>	2	6	5	4	5	2	9
<i>ODOI</i>	2	2	4	1	2	2	1
<i>BPBT</i>	13	25	27	11	12	12	29
<i>CFAT</i>	4	5	2	4	4	5	2

(continued)

Table 5.12 (continued)

Type gene	Pm	Md	Pt	At	Vv	Fv	Os
<i>BSMT</i>	21	32	25	23	25	15	13
<i>CCMT</i>	12	34	23	21	25	12	11
<i>BEAT</i>	34	16	17	3	4	14	–
<i>OOMT</i>	13	37	28	9	18	9	23
<i>IEMT</i>	2	44	30	15	12	14	13
<i>EGS</i>	9	13	18	8	18	10	7
<i>IGS</i>	2	10	17	8	17	10	7
<i>POMT</i>	6	43	32	15	12	14	11
<i>SAMT</i>	10	33	24	24	25	14	10
<i>PAAS</i>	6	4	5	2	5	6	7
<i>α-terpinene-synthase</i>	5	21	33	9	31	27	5
<i>β-pinene-synthase</i>	4	20	34	10	29	27	5
<i>Germacrene</i>	16	19	34	30	30	28	8
<i>TPS10</i>	8	13	13	10	7	9	8
<i>Linalool synthase</i>	1	2	2	1	–	1	2
<i>CCD</i>	6	12	16	7	7	6	5
<i>Limonene-3-hydroxylase</i>	68	111	108	93	50	56	97

SNP markers from eight linkages groups. We conclude that the combination of Illumina GA, whole-genome mapping technologies and the genetic map constructed by RAD can be used to perform de novo sequencing of plant genomes, allowing high-quality, rapid and low-cost sequencing of other plant species with similar conditions.

The sequence of the *P. mume* genome is a valuable resource for biological research and

breeding. Based on the sequences of *P. mume*, *M. \times domestica* and *F. vesca*, we reconstructed nine ancestral chromosomes of the Rosaceae family and inferred that they were shaped from an ancestor. Analysis of the *P. mume* genome and transcriptome can provide insights into the mechanisms of flowering scent, flowering dormancy and disease resistance. The genome also increases our knowledge of the evolution of the Rosaceae family and the function of the plant relative system.

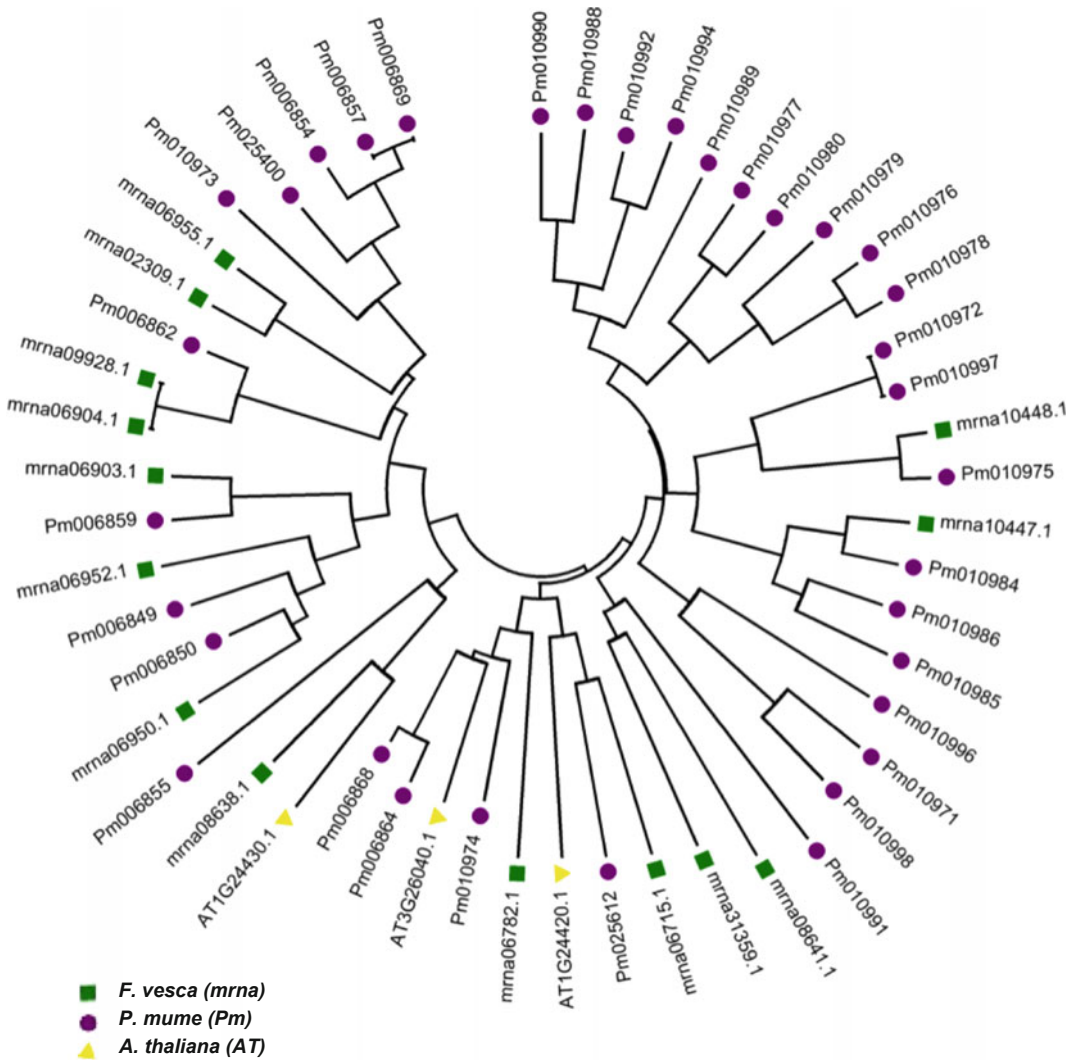


Fig. 5.11 Phylogenetic relationships of *BEAT* genes in *P. mume* (Pm), *F. vesca* (mrna) and *A. thaliana* (AT)

References

- Aranovich D, Lewinsohn E, Zaccari M (2007) Post-harvest enhancement of aroma in transgenic lisianthus (*Eustoma grandiflorum*) using the *Clarkia breweri benzyl alcohol acetyltransferase (BEAT)* gene. *Postharvest Biol Technol* 43(2):255–260. <https://doi.org/10.1016/j.postharvbio.2006.09.001>
- Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1):45–48. <https://www.ncbi.nlm.nih.gov/pubmed/10592178>
- Baxter SW, Davey JW, Johnston JS, Shelton AM, Heckel DG, Jiggins CD et al (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE* 6(4): e19315. <https://doi.org/10.1371/journal.pone.0019315>
- Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573–580. <https://www.ncbi.nlm.nih.gov/pubmed/9862982>
- Birney E, Clamp M, Durbin R (2004) GeneWise and genomewise. *Genome Res* 14(5):988–995. <https://doi.org/10.1101/gr.1865504>
- Chen J (1995) Some aspects on Chinese Mei flower research. *J Beijing For Univ* 17(S1):1–7
- Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, Cistue L, Corey A, Filichkina T et al (2011) Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in

- barley. *BMC Genom* 12(1):4. <https://doi.org/10.1186/1471-2164-12-4>
- Dudareva N, Raguso RA, Wang J, Ross JR, Pichersky E (1998) Floral scent production in *Clarkia breweri*. III. Enzymatic synthesis and emission of benzenoid esters. *Plant Physiol* 116(2):599–604. <https://www.ncbi.nlm.nih.gov/pubmed/9489012>
- Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM (2007) Creating a honey bee consensus gene set. *Genome Biol* 8(1):R13. <https://doi.org/10.1186/gb-2007-8-1-r13>
- Hampson SE, Gaut B, Baldi P (2004) Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* 21(8):1339–1348
- Horvath DP, Sung S, Kim D, Chao W, Anderson J (2010) Characterization, expression and function of *DORMANCY ASSOCIATED MADS-BOX* genes from leafy spurge. *Plant Mol Biol* 73(1–2):169–179. <https://doi.org/10.1007/s11103-009-9596-5>
- Hummer KE, Janick J (2009) Rosaceae: taxonomy, economic importance, genomics. In: *Genetics and genomics of Rosaceae*. Springer, pp 1–17
- Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161):463–477. <https://doi.org/10.1038/nature06148>
- Jiménez S, Lawton-Rauh AL, Reighard GL, Abbott AG, Bielenberg DG (2009) Phylogenetic analysis and molecular evolution of the dormancy associated MADS-box genes from peach. *BMC Plant Biol* 9(1):81
- Jiménez S, Reighard G, Bielenberg D (2010) Gene expression of *DAM5* and *DAM6* is suppressed by chilling temperatures and inversely correlated with bud break rate. *Plant Mol Biol* 73(1–2):157–167
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1–4):462–467
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30
- Kent WJ (2002) BLAT—the BLAST-like alignment tool. *Genome Res* 12(4):656–664
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z et al (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265–272
- Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955
- Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878–2879
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17(2):240–248
- Pichersky E, Dudareva N (2007) Scent engineering: toward the goal of controlling how flowers smell. *Trends Biotechnol* 25(3):105–110
- Salamov AA, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10(4):516–522
- Salse J, Abrouk M, Murat F, Quraishi UM, Feuillet C, (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief Bioinform* 10(6):619–630
- Sasaki R, Yamane H, Ooka T, Jotatsu H, Kitamura Y, Akagi T et al (2011) Functional and expression analyses of *PmDAM* genes associated with endodormancy in Japanese apricot. *Plant Physiol* 157(1):485–497
- Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL et al (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43(2):109
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids Res* 34 (suppl_2):W435–W439
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24(8):1596–1599
- Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 080978:108
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511
- Van Ooijen JW, Voorrips R (2001) JoinMap[®] 3.0, software for the calculation of genetic linkage maps. Plant Research International, Wageningen, Netherlands, pp 1–51
- Velasco R, Zharkikh A, Affourti J, Dhingra A, Cestaro A, Kalyanaraman A, et al (2010) The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat Genet* 42(10):833
- Wisniewski M, Norelli J, Bassett C, Artlip T, Macarasin D (2011) Ectopic expression of a novel peach (*Prunus persica*) *CBF* transcription factor in apple (*Malus × domestica*) results in short-day induced dormancy and increased cold hardiness. *Planta* 233(5):971–983
- Xing Z, Suo L, Liu H, Zhang Q-X (2009) The germplasm resources survey of *Prunus mume* at Tangmai. *Northem Hort* 10:46 (in Chinese)
- Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591
- Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848
- Zhou S, Kile A, Bechner M, Place M, Kvikstad E, Deng W et al (2004) Single-molecule approach to bacterial genomic comparisons via optical mapping. *J Bacteriol* 186(22):7773–7782. <https://doi.org/10.1128/JB.186.22.7773-7782.2004>