Novel Insights into Tree Biology and Genome Evolution as Revealed Through Genomics

David B. Neale,[1] Pedro J. Martínez-García,[1] Amanda R. De La Torre,[1] Sara Montanari,[1] and Xiao-Xin Wei[2]

[1]Department of Plant Sciences, University of California, Davis, California 95616; email: dbneale@ucdavis.edu

[2]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

Annu. Rev. Plant Biol. 2017. 68:457–83

First published online as a Review in Advance on February 6, 2017

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

https://doi.org/10.1146/annurev-arplant-042916-041049

Keywords

woody plants, genome size, transposable elements, perennialism, adaptation, fruit quality

Abstract

Reference genome sequences are the key to the discovery of genes and gene families that determine traits of interest. Recent progress in sequencing technologies has enabled a rapid increase in genome sequencing of tree species, allowing the dissection of complex characters of economic importance, such as fruit and wood quality and resistance to biotic and abiotic stresses. Although the number of reference genome sequences for trees lags behind those for other plant species, it is not too early to gain insight into the unique features that distinguish trees from nontree plants. Our review of the published data suggests that, although many gene families are conserved among herbaceous and tree species, some gene families, such as those involved in resistance to biotic and abiotic stresses and in the synthesis and transport of sugars, are often expanded in tree genomes. As the genomes of more tree species are sequenced, comparative genomics will further elucidate the complexity of tree genomes and how this relates to traits unique to trees.

Contents

INTRODUCTION

The fundamental importance and applied value of a reference genome sequence for any organism are widely recognized. In plants, the first reference genome sequence was obtained for the model plant *Arabidopsis thaliana* (5). The first tree genome to be sequenced, and just the third plant genome, was that of black cottonwood (*Populus trichocarpa* Torr. & Gray), which was published in 2006 (108); in the decade since then, another 40 tree genomes have been sequenced and published (**Table 1**). Several recent reviews have included genome sequencing of tree species as well as broader developments in tree genomics (17, 28, 87, 94).

The purpose of this review is to chronicle the history of reference genome sequencing in tree species by first reviewing the species sequenced and the technologies applied, and then focusing on three specific areas of plant biology where new knowledge has been gained that might not otherwise have been obtained from the reference sequences of nontree genomes. The three areas we discuss are (*a*) the noncoding and repetitive DNA content of tree versus nontree species, which might account for the large sizes of some tree genomes and the distinct characteristics of long-lived, perennial organisms; (*b*) genes, gene families, and expression patterns that underlie the perennial growth habit and biotic and abiotic adaptations to the environment; and (*c*) genes, gene families, and expression patterns that underlie edible fruit development and quality.

Eudicots (true dicots): a monophyletic clade that includes most of the dicot flowering plants; also called tricolpates

We have chosen to use a very broad definition of tree that includes perennial plants with an elongated stem as well as all seed plants (eudicots, monocots, and gymnosperms). We also

Table 1 Genomic resources in tree species

Family	Species	Common name	Genome size (Mb)	2*n*	Sequencing strategy	Sequencing technology	Assembly N50 scaffold (kb)	Assembly software	Reference(s)
Actinidiaceae	*Actinidia chinensis*	Kiwifruit	758	58	WGS	Illumina	646.8	ALLPATHS-LG and GapCloser	53
Arecaceae	*Elaeis guineensis*	African oil palm	1,800	32	WGS and BAC pool	Roche 454 and Sanger	1,270	Newbler (Roche GS De Novo Assembler)	102
	Phoenix dactylifera	Date palm	658[a]	36	WGS	Illumina	9.3 (excluding scaffolds of <500 base pairs)	SOAPdenovo	2
			671	36	WGS and BAC	Roche 454 and SOLiD	329.9	Newbler (Roche GS De Novo Assembler) and BioScope	3
Betulaceae	*Betula nana*	Dwarf birch	450	28	WGS	Illumina	18.7	SOAPdenovo-63mer	113
Caricaceae	*Carica papaya*	Papaya	372	18	WGS	Sanger	NA	Arachne	79
Euphorbiaceae	*Hevea brasiliensis*	Rubber tree	2,150[b]	36	WGS	Illumina, Roche 454, and SOLiD	3	Newbler (Roche GS De Novo Assembler)	97
			2,150[c]	36	WGS	Illumina and PacBio	67.2	Platanus and PBJelly2	66
			2,150[c]	36	WGS and BAC	Illumina	1,280	SOAPdenovo and SSPACE	105
	Jatropha curcas	Barbados nut	410	22	Pooled BAC	Sanger	3.8	PCAP.REP, MIRA	98
					WGS	Illumina and Roche 454			
Fagaceae	*Juglans regia*	Persian walnut	606	32	WGS	Illumina	465	SOAPdenovo and MaSuRCA	75
	Quercus robur	Pedunculate oak	740	24	WGS and BAC	Illumina, Roche 454, and Sanger	260	Newbler (Roche GS De Novo Assembler), SSPACE, and GapCloser	93
Malvaceae	*Theobroma cacao*	Cacao	430	20	WGS	Illumina, Roche 454, and Sanger	473.8	Newbler (Roche GS De Novo Assembler)	6
Meliaceae	*Azadirachta indica*	Neem tree	364[c]	30	WGS	Illumina, IonTorrent, and Sanger	452	SOAPdenovo	65
Moraceae	*Morus notabilis*	Mulberry	357[c]	14	WGS	Illumina	390.1	SOAPdenovo	47

(*Continued*)

Table 1 (*Continued*)

Family	Species	Common name	Genome size (Mb)	2*n*	Sequencing strategy	Sequencing technology	Assembly N50 scaffold (kb)	Assembly software	Reference(s)
Musaceae	*Musa acuminata*	Banana (A genome)	523	22	WGS (DH)	Roche 454 and Sanger	1,300	Newbler (Roche GS De Novo Assembler)	25
	Musa balbisiana	Banana (B genome)	438[d]	22	WGS	Illumina	467	CLC Genomics Workbench	27
Myrtaceae	*Eucalyptus camaldulensis*	Red river gum	650[e]	22	WGS and BAC	Roche 454 and Sanger	NA	CABOG	49
	Eucalyptus grandis	Flooded gum	640	22	WGS and BAC end	Sanger	L50 = 53,900[f]	Arachne and Rebuilder	81
Oleaceae	*Olea europaea*	Olive tree	1,400–1,500	46	WGS	Illumina and Roche 454	1.5	CLC Genomics Workbench and Minimus2 assembler	8
Pinaceae	*Picea abies*	Norway spruce	19,600	24	WGS (1*n*) and pooled fosmid (2*n*)	Illumina	4.9	CLC Assembly Cell, BESST, and GAM-NGS	85
	Picea glauca	White spruce	20,000[b]	24	WGS (2*n*)	Illumina	20.3	ABySS	13
			20,000[b]	24	WGS (2*n*)	Illumina	NG50 = 83.0[g]	ABySS	115
	Pinus lambertiana	Sugar pine	31,000	24	WGS (1*n* and 2*n*)	Illumina	246.6	SOAPdenovo and MaSuRCA	104
	Pinus taeda	Loblolly pine	22,000	24	WGS (1*n* and 2*n*)	Illumina	66.9	MaSuRCA	84, 116, 128
Rhamnaceae	*Ziziphus jujuba*	Jujube	444[c]	24	WGS and BAC-to-BAC	Illumina	301.1	SOAPdenovo and SSPACE	71
Rosaceae	*Malus* × *domestica*	Apple	750	34	WGS	Roche 454 and Sanger	L50 = 1.5[f]	Programs developed at Myriad Genetics	109
	Prunus mume	Mei	280[c]	16	WGS	Illumina	577.8	SOAPdenovo	126
	Prunus persica	Peach	265	16	WGS (DH)	Sanger	26.8	Arachne	111
	Pyrus communis	European pear	600	34	WGS	Roche 454	88.1	Newbler (Roche GS De Novo Assembler)	18
	Pyrus × *bretschneideri*	Chinese pear	527	34	WGS and BAC-to-BAC	Illumina	540.8	SOAPdenovo and SSPACE	119

Rubiaceae	*Coffea canephora*	Coffee	710	22	WGS (DH)	Illumina and Roche 454	1,261	Newbler (Roche GS De Novo Assembler) and GapCloser	30
					BAC	Sanger			
Rutaceae	*Citrus clementina*	Clementine mandarin	301[h]	18	WGS, fosmid, and BAC (1*n*)	Sanger	L50 = 31,400[f]	Arachne	118
	Citrus sinensis	Sweet orange	367	18	WGS	Roche 454 and Sanger	L50 = 250.5[f]	Newbler (Roche GS De Novo Assembler)	118
			367	18	WGS (DH)	Illumina	1,690	SOAPdenovo and Opera	120
Salicaceae	*Populus euphratica*	Desert poplar	593[c]	38	WGS and fosmid pool	Illumina	482	SOAPdenovo and SSPACE	74
	Populus trichocarpa	Black cottonwood	485 ± 10[c]	38	WGS and BAC	Sanger	3.1	Jazz	108
	Salix suchowensis	Purple willow	429	38	WGS	Illumina and Roche 454	925.0	Newbler (Roche GS De Novo Assembler) and SOAPdenovo	26
Thymelaeaceae	*Aquilaria agallocha*	Agarwood	736	16	WGS	Illumina	126.4	SGA, SSPACE, and GapFiller	20
Vitaceae	*Vitis vinifera*	Grapevine	475	38	WGS	Sanger	2,065	Arachne	56
			505[i]	38	WGS	Roche 454 and Sanger	NA	Programs developed at Myriad Genetics	110

This table includes all 41 published reference genome sequences for tree species as of October 2016. For each genome, it reports the estimated size (based on flow cytometry unless otherwise specified), diploid number, sequencing strategy and technology, scaffold N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found), software used for the assembly, and associated reference(s). Abbreviations: BAC, bacterial artificial chromosome; DH, double haploid; NA, not applicable; WGS, whole-genome sequencing.

[a]Based on the average of the flow cytometry analysis and calculation of the percentage of fully sequenced fosmids.

[b]Based on Feulgen microdensitometry.

[c]Based on *k*-mer depth distribution.

[d]Based on the assumption that the Pisang Klutuk Wulung haploid genome is 93.3% of the size of the double-haploid Pahang genome.

[e]Based on the average genome size of the subgenus *Symphyomyrtus*, which includes the majority of the most widely planted and bred species of *Eucalyptus*.

[f]L50 corresponds to N50.

[g]NG50 is identical to N50, except that the length of the genome being assembled is estimated as being equal to the average of the length of the two haplotypes.

[h]Size of the genome sequenced (no previous estimation).

[i]For the calculation, see Reference 110.

include one woody vine species, grapevine (*Vitis vinifera* L.), which has several biological and agronomical features in common with regular fruit tree crops. Key areas of plant biology that can yield new insights into plant biological processes and traits include the perennial growth habit, woody stems, and long generation times. Furthermore, because initial reference genome sequencing efforts have focused on species of interest to humans, the sequencing of tree genomes contributes to our knowledge of wood formation (forest trees), edible fruits (horticultural trees), or medicinal and industrial properties (oil and rubber trees).

Next-generation sequencing: modern sequencing technologies based on short-read massively parallel sequencing (such as Illumina/Solexa and Roche 454)

Tree Genomes Sequenced and Published to Date

As of October 2016, 41 tree reference genome sequences have been completed and published, comprising 35 species, 27 genera, and 20 families (**Table 1**). The phylogenetic distribution of these species reveals that most are eudicots (almost all rosids); only four are monocots and four are gymnosperms (**Figure 1**). This difference is undoubtedly due to the importance of tree fruits and seed oils from the eudicot group. The other profound difference between angiosperms and gymnosperms is the size of their genomes (**Table 1**). Angiosperm genomes that have been sequenced generally range in size from 0.5 to 1.0 Gb, whereas the gymnosperm (conifer) genomes are ~20 Gb or larger; until recently, high costs and technical limitations prohibited the sequencing of the latter. Even though only 35 of the few tens of thousands of known tree species have a reference genome sequence, the taxonomic representation is now broad enough that we can begin to understand what makes a tree a tree by comparing tree and nontree genomes.

Chronology and Sequencing and Assembly Strategies

The chronological order in which tree genomes have been sequenced follows a pattern of genome size, availability of funding, and importance to humans. The black cottonwood genome, the first tree genome to be sequenced (108), is relatively small (485 Mb), and although *Populus* is a globally valuable genus, it is much less important economically than conifers and *Eucalyptus*. Nevertheless, the US Department of Energy identified it as a potentially important lignocellulose energy species and provided funding through the Joint Genome Institute. The project was carried out with whole-genome sequencing of bacterial artificial chromosomes using Sanger technology and may have cost on the order of US$10 million. For the time, the resulting assembly and annotated genome were of very high quality. Interestingly, however, only two other Salicaceae genome sequences have been completed and published in the decade since then: those of the purple willow (*Salix suchowensis*) (26) and desert poplar (*Populus euphratica*) (74).

The next woody plant (although not a tree) to be sequenced was grapevine (*V. vinifera*); two different genome sequences were completed in 2007, both primarily using Sanger technology (56, 110). Following shortly thereafter, and also sequenced using Sanger technology, was the papaya (*Carica papaya* L.) genome (79). The apple (*Malus* × *domestica* Borkh.) genome sequence was completed in 2010 using mostly Sanger and some Roche 454 sequencing (109). Likewise, two *Eucalyptus* genomes were sequenced in this period using a combination of Sanger and next-generation sequencing: those of the red river gum (*Eucalyptus camaldulensis*) (49) and the flooded gum (*Eucalyptus grandis*) (81). The last genome to be sequenced using entirely Sanger technology was that of peach [*Prunus persica* (L.) Batsch], which was begun before 2010 and finally published in 2013 (111).

The transition from Sanger sequencing to next-generation sequencing started to take hold in about 2010, accelerating progress in tree genome sequencing. Nearly all tree genomes sequenced since 2010 were done by whole-genome sequencing using one or more next-generation-sequencing platforms (**Table 1**). However, the quality of the published genome assemblies varied

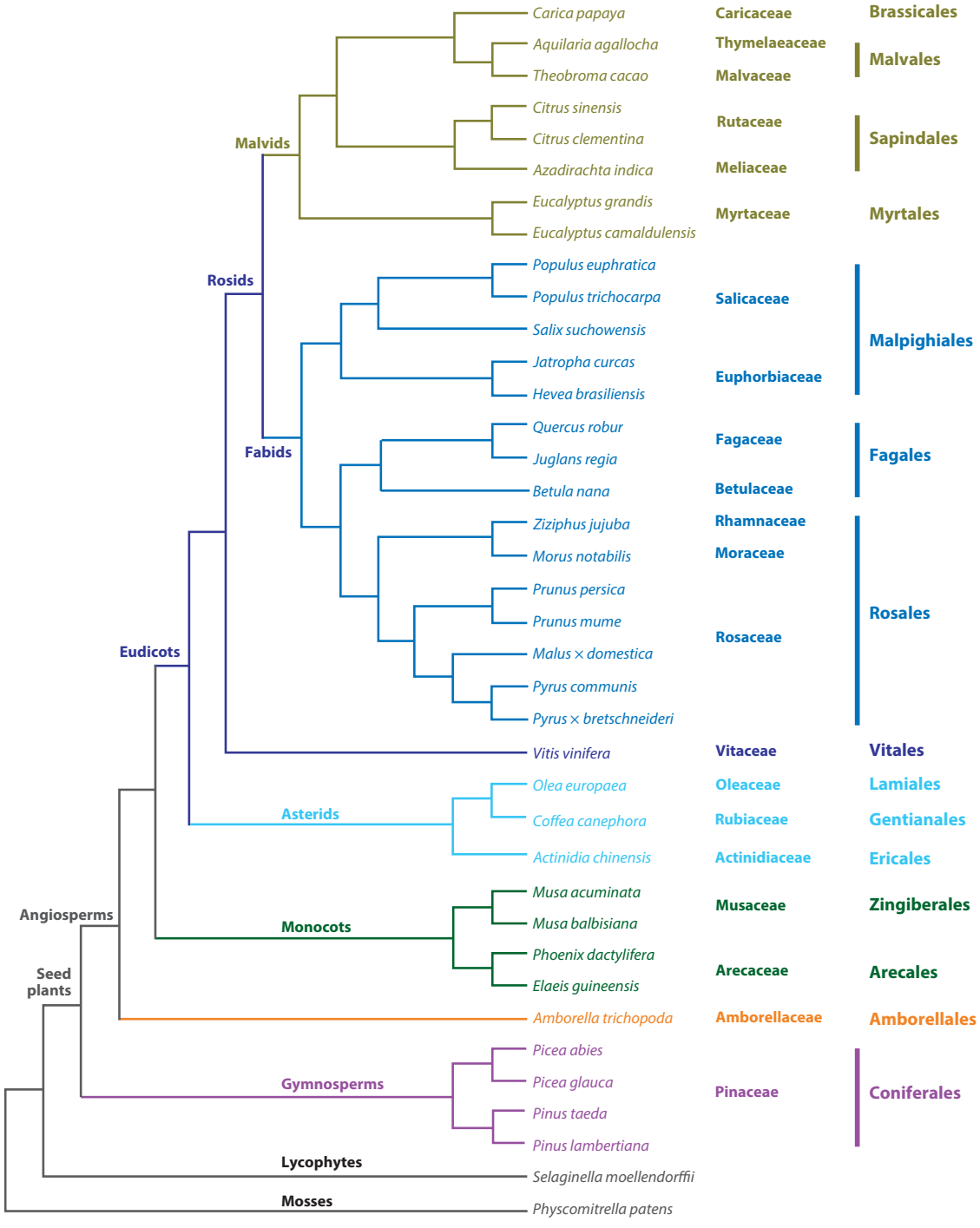

Figure 1

Phylogenetic tree of the 35 tree species with a published reference genome sequence, along with 1 additional angiosperm (*Amborella trichopoda*) and 2 nonseed plants as basal species (*Selaginella moellendorffii* and *Physcomitrella patens*). The overall topology is based on the Angiosperm Phylogeny Group (APG) IV classification system (4) and the green plants cladogram from the Tree of Life Web Project (78).

significantly (**Table 1**). The scaffold N50 sizes, which are indicative of the fragmented level of the genome, could vary by orders of magnitude, although a rough average for 2010–2016 would be a scaffold N50 of ~400 kb. During this time, many genome sequences of fruit trees were completed [banana (*Musa acuminata* and *Musa balbisiana*), cacao (*Theobroma cacao*), clementine mandarin (*Citrus clementina*), coffee (*Coffea canephora*), kiwifruit (*Actinidia chinensis*), olive (*Olea europaea*), pear (*Pyrus* × *bretschneideri* and *Pyrus communis*), sweet orange (*Citrus sinensis*), and Persian walnut (*Juglans regia*)] as well as the sequences of a few angiosperm forest trees [dwarf birch (*Betula nana*) and pedunculate oak (*Quercus robur*)], industrial-product trees [Barbados nut (*Jatropha curcas*), African oil palm (*Elaeis guineensis*), and rubber (*Hevea brasiliensis*)], and medicinal trees [agarwood (*Aquilaria agallocha*), jujube (*Ziziphus jujuba* Mill.), and neem (*Azadirachta indica*)].

The other major advance in tree genome sequencing that occurred with the transition to the next-generation-sequencing era was the sequencing of the very large genomes of a few conifer species (**Table 1**). Norway spruce (*Picea abies* L.) (85), white spruce [*Picea glauca* (Moench) Voss] (13), and loblolly pine (*Pinus taeda* L.) (84) were the first conifer genomes and the first gymnosperm genomes to be sequenced. The sequencing and assembly approaches differed among these projects, leading to large differences in the quality of the assemblies. One approach that was used for both the Norway spruce and loblolly pine projects was to prepare short-insert whole-genome-sequencing libraries from a single-haploid (1*n*) seed megagametophyte from the target tree. This avoided assembly challenges associated with the high heterozygosity of diploid tissues from trees.

As is the case with all types of organisms, tree reference genome sequencing and population-level resequencing are now proceeding rapidly. In this review, we have included only the 41 reference genomes that have been published, but an equal number, if not many more, either have been completed but are not yet published or are in progress. It is likely an underestimate to say that hundreds of tree genomes will be sequenced before the year 2020.

Scaffold N50: the scaffold size above which 50% of the total length of the sequence assembly can be found

Synteny: conservation of genomic regions within two sets of chromosomes

Whole-genome duplication (WGD): an event in which the entire genome of an organism is copied one or more times; sometimes called polyploidization

NONCODING DNA CONTENT AND GENOME SIZE VARIATION IN TREES AND THE EVOLUTIONARY MECHANISMS RESPONSIBLE FOR THESE DIFFERENCES

The variation and evolution of genome size in land plants are dynamic, with both increases and decreases (103), and in trees are just beginning to be understood (22, 70, 83). A great deal of synteny and collinearity has been found between tree genomes, even across different genera (43, 54, 89, 121). Comparisons of genomes and proteomes from sister species, such as willow and poplar or pear and apple, have confirmed their high collinearity (18, 26, 119). The tree genome blooming period (2010–2016) provided new tools that can be exploited to further understand the variation in genome size, the proportions of protein-coding sequences and highly repetitive noncoding DNA, and the relationship between these two. Two types of events, whole-genome duplication (WGD, also called polyploidization) and transposable element amplification, have been associated with rapid growth of genome size (also called genome obesity) in herbaceous angiosperms (11, 67). Many studies of insertions and deletions that affect genome size have used well-characterized and relatively small genomes of nontree plants, such as *Arabidopsis* (10, 31). However, similar studies in trees species are scarce, and the mechanisms controlling genome expansion or shrinkage in tree species in general, and in gymnosperms in particular, are not completely understood (68, 80).

Genome Size Variation

Both the smallest and the largest plant genomes ever found belong to herbaceous species (*Genlisea margaretae* and *Paris japonica*, respectively); these species have almost a 2,400-fold difference

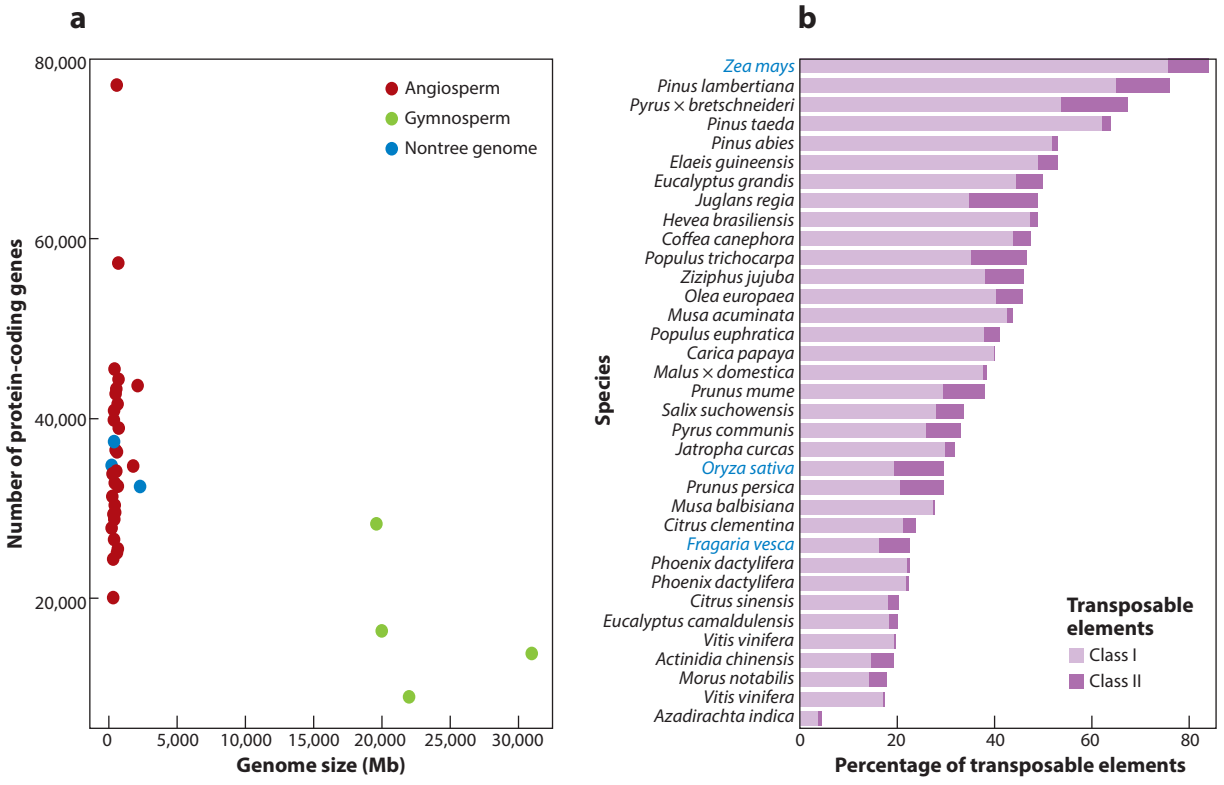

Figure 2

Genomic components of sequenced tree genomes. (*a*) Genome size and number of protein-coding genes in angiosperms and gymnosperms. The genomes of sequenced angiosperms are always smaller than those of gymnosperms, and genome size variation appears to be independent of the number of protein-coding genes. (*b*) Percentage of transposable elements in 30 of the 35 species with a published reference genome sequence, broken down into class I elements (retrotransposons, which transpose via an RNA intermediate) and class II elements (transposons, which move by a cut-and-paste mechanism). An increasing proportion of transposable elements is thought to be one of the main causes of positive unidirectional genome size variation in plants. Species listed twice (*Phoenix dactylifera* and *Vitis vinifera*) have two published genome sequences with differing proportions. The genomes of three nontree species (*Zea mays*, *Oryza sativa*, and *Fragaria vesca*, shown in *blue*) have been included for comparison.

in genome size (9, 90). The genome size variation in tree species is smaller, and among those with a published reference genome sequence (**Table 1**), the largest genome [belonging to sugar pine (*Pinus lambertiana*), 31 Gb] is only ~117 times the size of the smallest [belonging to peach (*P. persica*), 265 Mb]. Moreover, the genomes of sequenced angiosperm trees are always smaller than those of gymnosperms (**Figure 2**).

Another important observation is that genome size variation seems to be independent of gene content (protein-coding genes) and organism complexity (86), and the predicted gene numbers in sequenced trees confirm this hypothesis. Although 50,172 genes were initially found in *P. taeda*, only 9,024 were identified in an improved and less fragmented version of the genome sequence (84). On average, approximately 33,000 genes (not counting partial genes) have been annotated in tree genomes. Palazzo & Gregory (86) also emphasized that even two closely related species with similar biological characteristics and the same ploidy level can have significant differences in genome size, as is clearly the case for *P. taeda* and *P. lambertiana*. Furthermore, variation in genome size and variation in chromosome number are not correlated in flowering plants in general,

as underlined by Soltis et al. (103), and this is observed in angiosperm trees in particular as well (**Table 1**). We expect the same trend in gymnosperms, and indeed, there appears to be no such correlation in sequenced diploid species, all of which have $2n = 24$ chromosomes but different genome sizes (104).

One hypothesis that can partially explain genome size variation in trees is the presence of long introns. In pines, the maximum intron length observed was 891,919 base pairs for *P. taeda* (84) and 578,081 base pairs for *P. lambertiana* (104). In *P. taeda*, 6,267 (4.4%) of the introns were longer than 20 kb, which exceeds the intron lengths described in other tree species. As discussed by Nystedt et al. (85), these large introns could possibly have resulted from the insertion of multiple repetitive elements in these species.

Whole-Genome Duplication

WGD is a major evolutionary force in woody angiosperm genomes, as observed in black cottonwood, apple, banana, kiwifruit, and pear, all of which contain traces of at least two WGD events (25, 53, 108, 109, 119). By contrast, WGDs are rare in gymnosperm trees (28, 83), with the remarkable exception of the hexaploid coast redwood (*Sequoia sempervirens*) (99). However, a new study based on phylogenomic analyses of transcriptomes from 24 gymnosperms and 3 outgroups suggested that three ancient WGDs occurred during the evolution of gymnosperms (70), in contrast to the conclusions regarding the Norway spruce genome (85). Future versions of published genomes with improved contiguity will enable further analyses of polyploidy and genome evolution for these large genomes.

The complete genome sequences of angiosperm trees have provided information on polyploidy and genome evolution. The now widely accepted hypothesis of a hexaploidization of all eudicots (called the γ event) was first suggested by Jaillon et al. (56) when sequencing the *V. vinifera* genome, which has a chromosomal state that is highly similar to that of the paleohexaploid progenitor (with seven proto-chromosomes), and was later confirmed in apple (109), cacao (6), sweet orange (120), peach (111), mulberry (*Morus notabilis*) (47), and coffee (30). In addition, the genome of kiwifruit (*A. chinensis*) shows traces of two recent WGD events, referred to as Ad-α and Ad-β, which occurred ~26.7 and 72.9–101.4 Mya, respectively, after the divergence from tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum* L.). In the *Populus* lineage, the desert poplar (*P. euphratica*) and black cottonwood (*P. trichocarpa*) genomes shared at least two WGDs and exhibit extensive collinearity, with a divergence time of approximately 14 Mya (74). The rubber tree (*H. brasiliensis*) genome showed signs of the eurosid triplication (the γ event) and of a recent WGD that occurred ~15.3 Mya, before the burst of *Hevea* long-terminal-repeat insertion (105). Regarding monocots, the banana genome sequence helped to uncover the evolutionary history of the *Musa* lineage, which underwent three rounds of WGD followed by gene loss. These three WGDs are not shared by the Poales lineage, which went through an independent WGD event (25). Two genera of the Rosaceae tribe Pyreae, *Malus* and *Pyrus*, underwent a more recent WGD (30–45 Mya) followed by a speciation event (109, 119). An interesting case is that of mulberry, for which an extremely wide range of chromosome numbers across the genus *Morus* has been reported (14–308 chromosomes). These diverse levels of polyploidization might have enabled its rapid adaptation to different environments (47).

Noncoding DNA Content

The accumulation of transposable elements is thought to be one of the main causes of a positive unidirectional genome size variation in plants (12, 35). Nystedt et al. (85) proposed a model

in conifers in which these trees might have less transposable element removal than other organisms, which would then contribute to their large genome sizes. The published reference genome sequences, for example, suggest that the 10-Gb difference in the sizes of the sugar pine and loblolly pine genomes can indeed be explained by transposable elements, which account for ~25 Gb (79% of the genome, 67% being long-terminal-repeat retrotransposons) in sugar pine and ~16 Gb (74% of the genome) in loblolly pine. By contrast, the sequenced angiosperm trees exhibit nonlinear variation of the repetitive elements with respect to genome size, which highlights the well-documented idea that alternative mechanisms, in particular WGDs (12), can affect genome size variation in this group. Retrotransposon content varies across the 35 sequenced tree species (**Figure 2**) but is always higher than the DNA transposon content. Ty3/Gypsy and Ty1/Copia long terminal repeats are the most abundant transposable elements in these genomes; however, no specific trend is evident in the content of these elements across tree genomes. Future improvements of the fragmented tree genome sequences will help clarify this complex landscape of noncoding DNA elements and their impact on the genome evolution of trees.

Long-terminal-repeat retrotransposons: retrotransposons that are similar in structure and life cycle to retroviruses and are abundant elements in eukaryotic genomes

Retrotransposons: DNA sequences that transpose via an RNA intermediate (replicative mechanism); they fall into two main groups depending on whether long terminal repeats flank the retroelement main body

Transposons: DNA sequences that move from one genomic location to another by a cut-and-paste mechanism; DNA transposons consist of a transposase gene flanked by two terminal inverted repeats

Identification of Novel Classes of Noncoding RNAs

The characterization of noncoding RNAs, such as small RNAs and long noncoding RNAs, has shed new light on the regulation of gene expression in trees. Small RNAs are associated with epigenetic processes and control of repetitive element proliferation (76) and can be classified as small interfering RNAs, microRNAs, and Piwi-associated RNAs (16). Long noncoding RNAs participate in various cellular processes, including mRNA splicing and ribosome biogenesis (96). The characterization of microRNAs in the grapevine genome suggested that this species has a complex RNA processing machinery, with four Dicer-like (DCL) proteins and nine *Argonaute* genes identified (110). In addition, 56 RNA-dependent DNA polymerase genes were potentially targeted by two microRNA gene families, miR396 and miR846—an unprecedented observation for a plant species. In cacao (*T. cacao*), most of the 91 predicted microRNAs putatively target mRNAs that encode transcription factors, which suggests a role as major regulators of gene expression (6). Sequencing of the banana species *M. balbisiana* seemed to indicate that 18 predicted novel microRNA families are B genome–specific in function and have evolved after the divergence from *M. acuminata* ~4.6 Mya (27). *P. trichocarpa* underwent a 1.9× expansion in microRNA genes in comparison with *Arabidopsis*, primarily in the miR169 and miR159/319 families (108). In *P. euphratica*, the novel microRNAs (119) were extensively up- or downregulated in response to salt stress (74). In mulberry, the predicted microRNA genes were associated with plant–herbivore interactions (with the silkworm, *Bombyx mori*) at a molecular level (47).

In gymnosperms, 13,031 spruce-specific intergenic long noncoding RNAs were annotated in the *P. abies* genome. These loci contain fewer exons, are shorter, and have a more tissue-specific expression than protein-coding loci (85). Moreover, two classes of small RNAs, the 24-nucleotide (nt) short RNAs (which are rarely expressed in conifer genomes) and the 21-nt short RNAs, were also identified. The 24-nt short RNAs, which are highly specific to reproductive tissues, are involved in silencing transposable elements through the establishment of DNA methylation; the 21-nt short RNAs are associated with genes, repeats, and promoters or untranslated regions. Recent work in *P. lambertiana* has confirmed the existence of a 24-nt DCL3 pathway in conifers, albeit with distinct spatial and/or temporal characteristics (40). Conifer genome sequences have enabled an unprecedented survey of the noncoding RNA landscape in gymnosperms that will shed light on the mechanisms for controlling tree genome obesity.

GENES, GENE FAMILIES, AND EXPRESSION PATTERNS THAT UNDERLIE THE PERENNIAL HABIT AND ADAPTATION TO THE ENVIRONMENT IN TREES

The Perennial Habit: Genes Associated with the Floral Transition, Bud Dormancy, and Woody Growth

Trees differ from herbaceous annuals in many ways. Most trees display a perennial growth behavior characterized by a multiple-year delay in flowering and, in temperate or boreal regions, an annual cycling between growth and dormancy; annual plants by contrast, grow, reproduce, and senesce within a single growing season. As asked by Groover (44): What genes make a tree a tree, as opposed to an annual plant? The availability of the *P. trichocarpa* genome (108) opened a window to address the tree-specific questions and expanded our understanding of the genetic mechanisms of plant adaptation to environmental change.

In plants, annualism and perennialism are two major reproductive strategies. Flowering time is a key factor in a plant's ability to adapt to the environment and optimize yield. Although extensive studies have revealed the molecular basis of the floral transition in the annual plant *Arabidopsis*, the genetic mechanisms that control these phases in trees were unknown until recently. Given the extended delay in flowering in trees, we may expect the function of the flowering-time genes to differ between annual plants and trees.

Interestingly, using the information contained in the *P. trichocarpa* genome, Böhlenius et al. (14) and Hsu et al. (52) examined the specific genes that regulate the juvenile-to-adult phase transition and floral transition and showed that the *CONSTANS/FLOWERING LOCUS T* (*CO/FT*) regulatory module that controls flowering time in annual plants also controls flowering in perennial aspen trees (*Populus* spp.), implying that the function of these flowering-time genes could be conserved between annual plants and trees. But unexpectedly, the authors also showed that the *CO/FT* module also controls short-day-induced growth cessation and bud set in the fall. In a subsequent study, Hsu et al. (51) further revealed that the *FT1* and *FT2* genes, which resulted from a whole-genome salicoid duplication event (108), coordinate the repeated cycles of vegetative and reproductive growth in woody perennial poplar. *FT1* determines reproductive onset in response to winter temperatures, and *FT2* promotes vegetative growth and inhibition of bud set in response to warm temperatures and long days in the growing season.

Given that annuals generally have a perennial ancestor (36), an important question is to what extent the functions of the flowering-time genes are conserved between trees in different systematic groups, particularly in gymnosperms, which first appeared 300 Mya. The extant gymnosperms include four lineages—conifers, cycads, *Ginkgo*, and gnetophytes—and all species are perennial. The *FT* gene encodes a small protein of approximately 175 amino acids that belongs to the phosphatidylethanolamine-binding protein (PEBP) family. Phylogenetic analysis divided the *PEBP* gene family into three subfamilies: *FT*, *TERMINAL FLOWER 1* (*TFL1*), and *MOTHER OF FT AND TFL1* (*MFT*) (19, 48). Studies have suggested that *FT* exists only in angiosperms and that gymnosperms lack it, instead containing only a group of *FT/TFL1*-like genes (45, 58, 62).

Gymnosperms and angiosperms are the two major groups of seed plants. They differ in many ways, particularly in their reproductive development and the strategies of adaptation and evolution. Because of their earlier appearance, gymnosperms play a key role in understanding the origin of angiosperms and unraveling the basis of some important evolutionary innovations in seed plants. Therefore, investigations of gymnosperm genomes can provide new insights into the functional evolution of flowering-time genes and reproductive evolution of seed plants.

The recently decoded gymnosperm genomes provide a powerful platform to study these genes. In Norway spruce (*P. abies*), six *FT/TFL1*-like genes (*PaFTL1–PaFTL6*) were found, four of which were newly identified (85). Taking advantage of all conifer genome sequences, Liu et al. (72) first confirmed that even gymnosperms have orthologous *FT*-like genes. The two previously described Norway spruce *FT/TFL1* genes, *PaFTL1* and *PaFTL2*, are real *FT* homologs, whereas the four newly identified genes in the Norway spruce genome are *TFL1*-like genes. The authors also found more *FT*-like and *TFL1*-like genes in *P. taeda* and *P. lambertiana* genomes that resulted from an additional gene duplication in pines: These two pines each have 3 *FT*-like genes and have 11 and 6 *TFL1*-like genes, respectively. In Norway spruce and Scots pine (*Pinus sylvestris* L.), *FTL2* plays a crucial role in bud set and growth cessation (7, 45, 57, 58). The tight relationship between the conifer *FTL2* gene and growth rhythm was further confirmed by expression studies of the Armand pine (*Pinus armandii*) *FTL2* gene (72). Another interesting finding in the study by Liu et al. (72) is that *FTL2* also has a potential role in maintaining female cone development in the pine family. By contrast, the *Ginkgo biloba FTL1* gene is suspected to be involved in bud dormancy (72).

Whole-genome scan: a type of study that uses whole-genome data to identify genes under diversifying selection

Diversifying selection: selection of alleles that confer a fitness advantage to individuals or populations

Woody growth is another feature of trees. Comparative analysis of the *P. trichocarpa* and *Arabidopsis* genomes showed that the genes responsible for cambium function and woody growth are not unique to woody plants; they are also expressed in the regulation of the shoot apical meristem in *Arabidopsis* (108). Papaya is a perennial that accumulates lignin in the cell wall at an intermediate level between *Arabidopsis* and poplar; the papaya genome (79) therefore provides a stepping-stone between woody and herbaceous plants. Genomic analysis has consistently shown that papaya has intermediate numbers of lignin synthetic genes (32)—fewer than poplar (39) but more than *Arabidopsis* (20). It has also revealed that, despite a closer evolutionary relationship to *Arabidopsis*, papaya shares with poplar an increased number of genes associated with cell expansion and lignin biosynthesis, possibly because of its larger plant size and the convergent evolution of the tree-like habit (79). A comparison of *Hevea* and *Populus* genomes revealed a noticeable difference in the numbers of caffeic acid *O*-methyltransferase (COMT) and cinnamyl alcohol dehydrogenase (CAD) proteins (10 COMTs and 5 CADs in *Hevea*, compared with 41 COMTs and 24 CADs in *Populus*) (97), which was suspected to be associated with the hardness of poplar wood compared with rubber wood. However, the semiwoody herb papaya has a large number of CADs (20) and only a single COMT (79). These differences imply that there is still much to learn about the genetic basis of what makes a tree a tree. As the amount of genomic data increases, comparative genomic analysis employing broad taxon sampling will be necessary to understand the details of some specific functional genes.

Adaptation to the Environment: Abiotic Stress

The ability of trees to respond to environmental signals by synchronizing growth and development with seasonal changes determines their distribution, potential for adaptation, and productivity (33). Current and predicted environmental changes have raised concerns about how tree species and populations will adapt to new environmental conditions, including prolonged droughts, increased salinization of soil and water, and cold temperature episodes (46, 88). Reference tree genome sequences have opened a window to understand the genomic architecture of adaptation to the environment and the complex biological processes that underlie responses to biotic and abiotic stress (**Figure 3**). Whole-genome scans in *Populus* spp. and comparative genomics studies in *Picea* spp. have shown that genes under diversifying selection often show overrepresentation of genes underlying responses to biotic and abiotic stress (29, 34, 37, 50). Similarly, recent expansions and subsequent subfunctionalization of gene families involved in biotic and abiotic stress [e.g.,

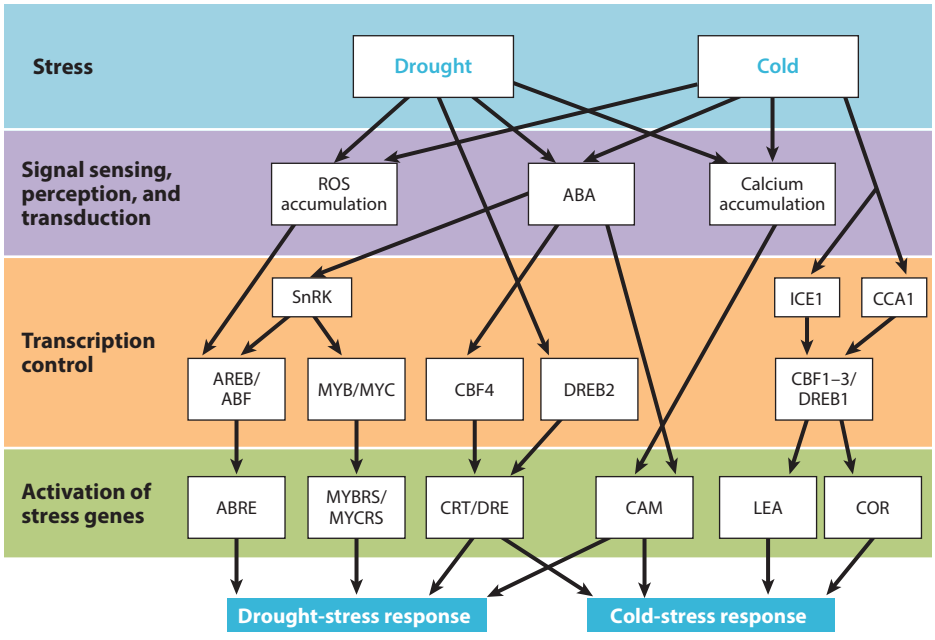

Figure 3

Simplified model of abiotic stress responses and signaling pathways leading to stress responses. Abbreviations: ABA, abscisic acid; ABF, ABRE-binding factor; ABRE, ABA-responsive element; AREB, ABRE-binding protein; CAM, calmodulin; CBF, CRT-binding factor; CCA1, circadian clock–associated 1; COR, cold responsive; CRT, C-repeat; DRE, dehydration-responsive element; DREB, DRE-binding protein; ICE1, inducer of CBF expression 1; LEA, late embryogenesis abundant; MYB, plant homolog of vertebrate MYB oncogene; MYBRS, MYB-recognition site; MYC, plant homolog of vertebrate MYC oncogene; MYCRS, MYC-recognition site; ROS, reactive oxygen species; SnRK, sucrose nonfermenting 1–related kinase. Figure modified from Reference 33.

pathogenesis-related 1 (PR1), leucine-rich repeats (LRRs), and abscisic acid/water deficit stress (ABA/WDS)] may have contributed to the adaptive radiation in *Eucalyptus* (60).

The sequencing of *P. trichocarpa* and, more recently, of *P. euphratica* (74, 108) has led to a growing understanding of *Populus* responses to drought stress, especially differential expression analyses using RNA sequencing or microarrays (24, 101, 106, 107, 124, 125). In drought-stressed *P. trichocarpa*, photosynthetic rates and water transportation were significantly reduced and accompanied by a repressed expression of a large number of genes involved in photosynthesis and cell development, leading to a strong reduction of tree growth (106). This response is distinctly different from that of its drought-adapted congener *P. euphratica*, which can maintain normal growth in dry and saline soils (107).

P. euphratica, a desert species, has gone through expansion of gene families associated with salt stress in comparison with its mesophytic congener *P. trichocarpa*. Several expanded gene families—including the *Myb*, *ethylene-responsive element binding factor* (*ERF*), *basic leucine zipper* (*bZIP*), and *WRKY* families—are involved in the transcription control of drought- and salt-stress responses. In addition, differentially expressed genes showing evidence of diversifying selection and recent gene family expansions are significantly enriched in functional categories related to salt stress, such as ion transport, ATPase activity, and oxidoreductase activity (74). Increased salt tolerance in *P. euphratica* may have developed through gene duplication and/or upregulation

of multiple genes involved in ion transport and homeostasis [e.g., *high-affinity K^+ transporter 1* (*HKT1*), *NhaD-type Na^+/H^+ antiporter* (*PeNhaD1*), *K^+ uptake transporter* (*KUP3*), and *Na^+/Ca^{2+} exchanger-like protein* (*NCL*)]. Similarly to *P. euphratica*, the jujube (*Z. jujuba*) medicinal tree is well adapted to drought and salinity in the arid regions of Asia. High levels of gene expression at all stages of fruit development have been found in species-specific genes involved in osmotic stress (71). The C-repeat-binding factor (CBF) and dehydration-responsive element–binding protein 2 (DREB2) transcription factors (discussed more below) and the ABA/WDS protein were found to be important players in adaptation to heat and drought in the flooded gum (*E. grandis*) (15, 60).

Cold stress and cold hardiness have been extensively studied in tree species. In fruit trees, the motivation was to understand the effects of frost episodes and cold storage of fruits for export (33); in forest trees, the motivation was to predict the performance of genotypes in breeding programs and to understand the ecology and physiology of cold-stress responses. The primary regulators of cold responses, and the best-characterized such regulators in model plants, are CBF and DREB1 (46). These transcription factors and their regulators, such as inducer of CBF expression 1 (ICE1), play a key role in initiating cold acclimation and inducing the expression of a large number of genes involved in cold responses, such as *cold responsive* (*COR*) and *late embryogenesis abundant* (*LEA*) genes. Recent studies of CBF/DREB have shown a different and more complex regulation and impact on abiotic-stress resistance in trees than in herbaceous plants (117). An increased number of CBF-binding sites in dormancy-associated MADS box (DAM) transcription factors may explain the early dormancy release and early spring flowering of the cold-tolerant mei tree (*Prunus mume*) (126).

While performing genome-wide analysis and expression profiling of the DREB transcription factor family in apple (*M. × domestica*), Zhao et al. (127) found that DREB-encoding genes are upregulated under various abiotic-stress treatments, suggesting that they play an important role in stress adaptation in *Malus*. More recently, a genome-wide study of the APETALA 2 (AP2)/ERF transcription factor family suggested a dramatic recent expansion of the CBF/DREB1 subfamily in *E. grandis*, which has the largest number of CBF/DREB1 genes ever reported for a dicot plant (15). Interestingly, the 17 *CBF* genes found in *E. grandis* showed differential expression in different tissues and under different applied stresses, suggesting that they have complementary rather than redundant functions. The large number of *CBF* genes may be the key for the winter survival of this tree, which originated in a tropical or subtropical climate and is now grown in more temperate climates (15, 82). Similarly, overexpression of CBF1/DREB1 in ecodormant buds of pedunculate oak (*Quercus robur*) and sessile oak (*Quercus petraea*) increased the tolerance of meristematic cells to cold temperatures (69). In addition to CBF/DREB, the WRKY and NAC transcription factors also play important roles in the transcriptional regulation of early cold response in *P. euphratica* (21). Dehydrins, a LEA subfamily, also play an important role in cold adaptation in tree species, and some cold-inducible dehydrins seem to be regulated by CBF transcription factors in several tree species, including *Quercus* spp. (46, 69).

DREB2: dehydration-responsive element–binding protein 2

ICE1: inducer of CBF expression 1

LEA: late embryogenesis abundant

MADS box: a gene family coding for transcription factors that in plants control flowering and fruit development and ripening

NBS-LRR: nucleotide-binding site–leucine-rich repeat

Biotic Stress

In contrast to annual plants, woody perennials need to develop long-term defense strategies to respond to biotic stress. Genes encoding nucleotide-binding site–leucine-rich repeat (NBS-LRR) proteins make up one of the largest gene families in plants and the largest class of disease resistance genes. Two main NBS-LRR classes may be present in plants depending on whether they have a coiled-coil motif [CC-NBS-LRR (CNL) class] or a Toll or interleukin-1 receptor domain [TIR-NBS-LRR (TNL) class] at the N terminus (32). Reference tree genome sequences and subsequent studies have revealed an extensive and considerably larger NBS-LRR gene family in tree species (such as *E. grandis*, *H. brasiliensis*, *M. × domestica*, *P. abies*, *P. taeda*, *P. trichocarpa*, *P. mume*,

Climacteric: characterized by a peak in respiration and a burst of ethylene at the onset of fruit ripening

P. persica, *V. vinifera*, and *Z. jujuba*) than in herbaceous plants (such as *Arabidopsis* and *S. lycopersicum*) (23, 64, 71, 108, 109, 123, 126). Kiwifruit (*A. chinensis*), a tree with a small number of NBS-LRR genes, and rice (*Oryza sativa*), an herbaceous monocot with a large number of non-TIR-NBS-LRR genes, seem to be the exception to the pattern observed (53, 122). Tandem duplications have been linked to recent expansions of the NBS-LRR family in species such as *E. grandis*, *M.* × *domestica*, *P. trichocarpa*, *P.* × *bretschneideri*, and *V. vinifera* (23, 91, 119, 123). In *E. grandis*, the cluster or supercluster arrangement of NBS-LRRs was correlated with differential expression responses to the pathogens *Chrysoporthe austroafricana* and *Leptocybe invasa*, suggesting functional relevance for the physical arrangement of the gene family (23).

Regions hosting resistance genes evolve rapidly, which could explain the unusually high nucleotide diversity in a genome region with a high density of genes encoding NBS-LRR proteins in *P. persica* (111). The CNL class is often more abundant than the TNL class in tree species; however, a larger number of TNL-class genes have been observed in species such as *E. grandis*, *P. abies*, and *P. taeda*, in a pattern similar to that of *Arabidopsis* (23, 84). It has been suggested, however, that distinct TNLs have expanded in conifers and angiosperms (84). In addition, TNL-class genes have been identified as strong candidates for the fusiform rust resistance gene *Fr1* in *P. taeda* and the white pine blister rust resistance genes *Cr1* and *Cr2* in *P. lambertiana* and western white pine (*Pinus monticola*) (84, 104).

GENES, GENE FAMILIES, AND EXPRESSION PATTERNS THAT UNDERLIE FRUIT DEVELOPMENT AND FRUIT QUALITY

Angiosperm species produce a wide variety of fruits, which can be distinguished by their type (whether they are dry or fleshy, and whether they are achenes, nuts, berries, pomes, or drupes) or ripening physiology (climacteric or nonclimacteric), although these categories are not phylogenetically constrained (100) (**Figure 4**). Fruit chemistry (the amount and composition of metabolites that affect the flavor and nutritional content of fruits) is complex (63) and has been strongly shaped by domestication; it is therefore difficult to find correlations with the subdivisions within angiosperms clades. In particular, none of the fruit characteristics seem to be associated with tree status.

Gene predictions and annotations of reference genomes, as well as their comparison with previously published genomes from different plants, have enabled the identification of gene families that are expanded or specific for certain groups of species. Here, we report on the main gene families related to fruit development and ripening, sugar metabolism, and flavor, as well as those involved in the biosynthesis of specific components considered particularly healthy and beneficial (such as vitamin C).

Fruit Development and Ripening

A review by Seymour et al. (100) described the complicated hormonal and genetic regulation of fruit development and ripening, encompassing all of the main transcription factors associated with these stages. The large family of MADS-box genes is present in all eukaryotic genomes analyzed so far, but the largest numbers of these genes are present in angiosperm species, where a considerable expansion occurred via gene duplications (1). Although the expansion of type II MADS-box genes was attributed mainly to WGDs, type I genes seem to have expanded through tandem duplications.

In plants, MADS-box genes control all major aspects of floral and fruit development. Type II MADS-box genes are well studied and are known to be involved mainly in floral organogenesis,

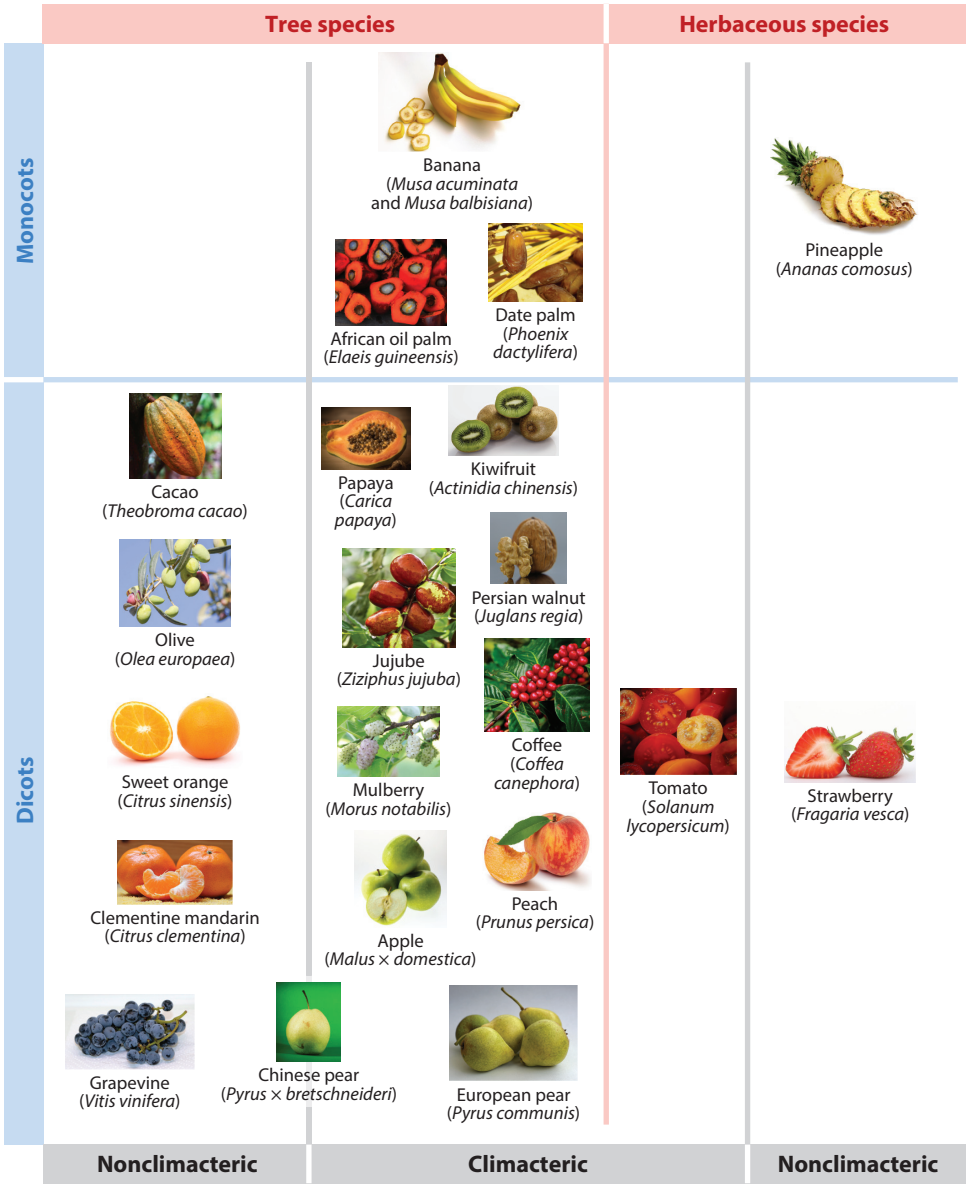

Figure 4

Edible fruits from 19 tree species plus those from 3 herbaceous species for comparison. Species are divided based on whether they are woody or herbaceous, monocots or dicots, and climacteric or nonclimacteric. Only fruits from species with a published reference genome sequence are included. Chinese pear (and Asian pears in general) can be climacteric or nonclimacteric depending on the variety. All images are from Pixabay (**http://pixabay.com**) and are free of copyright under Creative Commons CC0 Public Domain.

regulation of flowering time, and fruit formation; type I MADS-box genes are not completely understood, but the few that have been characterized have roles in gametophyte, embryo, and seed development (41). The papaya genome has many fewer genes than other sequenced genomes but has more type I MADS-box genes than *Arabidopsis* (which is also a member of the Brassicales) (79). This was confirmed more recently by Gramzow & Theißen (42), who annotated MADS-box genes in 17 plant genomes and discovered more than 2,000 such genes; they reported that the papaya genome has 262 MADS-box genes, of which 229 are type I, the largest number among all 17 studied species. The large difference between the numbers of type I and type II MADS-box genes in this species is consistent with the absence of further WGDs after the γ event in this genome. However, papaya's large overall number of MADS-box genes—even larger than the number in tomato, which has 190—still requires an explanation. Among the 17 genomes analyzed by Gramzow & Theißen (42), the *M.* × *domestica* genome had the largest number of type II genes, 15 more than *Arabidopsis* (118). Velasco et al. (109) hypothesized that the expansion of MADS-box genes in the *M.* × *domestica* genome may be related to the formation of the pome, which is a peculiarity of the Pyreae tribe.

MADS–RIPENING INHIBITOR (MADS-RIN): a transcription factor that induces increases in respiration and ethylene at ripening; it was first discovered in tomato

Colorless nonripening locus containing a SQUAMOSA promoter binding protein (SBP)–like box (CNR-SBP): a protein that plays a major role in fruit ripening

The MADS–RIPENING INHIBITOR (MADS-RIN) and colorless nonripening locus containing a SQUAMOSA promoter binding protein (SBP)–like box (CNR-SBP) proteins have central roles in the regulation of ripening in both climacteric and nonclimacteric fruits, as first suggested by studies of tomato and strawberry (*Fragaria vesca*) (39, 112). This hypothesis was confirmed by the sequencing of fruit tree species, which enabled, for example, the identification of *RIN*-like genes controlling ripening in apple (55) and kiwifruit (77), the observation of upregulated MADS-RIN in ripening oranges (120), and the discovery of expressed SBP-box genes in grapevine fruit (114). The identification of genes related to ethylene signaling in both climacteric and nonclimacteric fruits, such as those encoding AP2/ERF or ETHYLENE RESPONSE ELEMENT BINDING PROTEIN (EREBP) in the banana species *M. acuminata* (25) and grapevine (*V. vinifera*) (110) and ETHYLENE RECEPTOR (ETR) in sweet orange (*Citrus sinensis*) (120) and grapevine (*V. vinifera*) (110), further supported the hypothesis of a common ripening mechanism for all fleshy fruits (38, 39, 59).

The Metabolism of Sugars

The composition and content of sugars have an important influence on the quality of most fruits. Annotation of the apple, banana, jujube, papaya, peach, pear, and date palm (*Phoenix dactylifera*) genomes has shown that gene families related to the synthesis and transport of sugars are often expanded in these species. For example, Ming et al. (79) observed that, although papaya has fewer predicted genes than *Arabidopsis*, it has more starch-associated genes, probably because of its greater need for storage in leaves, stems, and developing fruit. RNA-sequencing analysis performed in mature green banana fruits after treatment with ethylene showed that three starch synthase genes were downregulated and one β-amylase gene was upregulated (25). Interestingly, this analysis also suggested that WGD may have caused subfunctionalization of two paralogous invertase genes, with potentially important consequences for the balance among sucrose, glucose, and fructose in ripened bananas. Dates have a high sugar content when ripened, and indeed, genes involved in the carbohydrate metabolism are much more upregulated than most of the other molecular events at the later stages of fruit development (3). Moreover, by examining the expression of sugar metabolism genes at different stages of fruit development in *P. dactylifera*, Al-Mssallem et al. (3) observed that enhanced accumulation of sugars in dates may result from a carbon fixation or refixation reaction. Jujube is another fruit with an extremely high sugar content (25–30%, twice as high as the content of most common fruits) (71). Compared with other sequenced species of the

order Rosales, genes involved in the metabolism of starch, sucrose, galactose, fructose, mannose, nucleotide sugar, and amino sugar are expanded in jujube. Moreover, most genes related to the major facilitator superfamily sugar transporter are overexpressed at ripening.

A peculiarity of most Rosaceae species [all Dryadoideae and Spiraeoideae but not Rosoideae (95)] is that photosynthetic carbohydrates are produced and translocated in the phloem mainly as sorbitol, which also accumulates in the fruit (73). Indeed, compared with other plant genomes, the genomes of apple, peach, and pear have considerably more copies of key genes related to the sorbitol metabolism (109, 111, 119). Apple, for example, has 71 sorbitol metabolism genes, whereas in other species (*Arabidopsis*, *Brachypodium*, cucumber, grapevine, maize, poplar, rice, sorghum, and soybean) the number ranges between 9 and 43 (109).

Sorbitol: the main photosynthetic product and phloem-translocated carbohydrate in the majority of the Rosaceae

Terpenes: a large class of compounds that play an important role in plant defenses and are responsible for the aroma of many fruits

Ascorbic acid (vitamin C): an essential nutrient with a main antioxidant function synthesized by several fruit plants

Flavor: Key Features for Fruit Quality

Edible fruits are destined for the fresh market or transformed (for example, turning grapes into wine), or their seeds can be used for the extraction of oils. Regardless of the final product, flavor—which is determined by aroma, taste, texture, and color (61)—is always an important aspect of fruit quality. Terpenoids are secondary metabolites responsible for the aroma of both wine and cocoa, and their synthesis is driven by terpene synthases. A total of 89 functional genes coding for terpene synthases and 27 pseudogenes were identified in the grapevine genome (56), and 57 and 9 (respectively) were identified in the cacao genome (6)—more than the 30–40 genes found in *Arabidopsis*, rice, and poplar. In cacao, the terpene synthase families of linalool synthase (a monoterpene) and cadinene synthase (a sesquiterpene) are particularly expanded, whereas in grapevine, the monoterpene synthases are highly diversified (56, 110).

In pear, the metabolism of α-linolenic acid is likely to be important for the fruit's aroma. The key enzymes are lipoxygenase and alcohol dehydrogenase, whose encoding genes are expanded in both pear and apple and are highly expressed during fruit development (119). Aroma in coffee beans results mainly from linoleic acid, a different fatty acid, which also contributes to flavor retention after roasting; the *C. canephora* genome contains six genes for the oleate desaturase fatty acid desaturase 2 (FAD2) (compared with one gene in *Arabidopsis*), the enzyme responsible for the synthesis of linoleic acid, which were probably generated by tandem duplications (30). Moreover, in the *C. canephora* genome, genes coding for *N*-methyltransferases involved in the synthesis of caffeine are expanded compared with the genomes of *Arabidopsis*, grapevine, and tomato, and they might have also originated by tandem duplications independently of cacao and tea plant (*Camellia sinensis*), two of the few other plants capable of synthesizing caffeine (30).

Fruit texture contributes considerably to quality in pome fruits. A characteristic feature of pears is the presence of stone or grit cells in the flesh, which is very rare in other fruits. The primary component of stone cells is lignin, and in the Chinese pear (*P.* × *bretschneideri*) genome, Wu et al. (119) observed expansion of gene families related to lignin synthesis. When analyzing the genome of the European pear (*P. communis*), Chagné et al. (18) focused on the expression of expansin genes, which are cell wall–related genes that influence fruit softening at ripening and may be linked to the melting texture typical of European pears. Indeed, although genes from this family were not expanded in European pear compared with crispier apples, their expression levels were higher.

Fruit Characteristics Associated with Beneficial Effects in the Human Diet

An important characteristic common to multiple fruits and vegetables is their high content of ascorbic acid (vitamin C), which contributes to their high nutritional value. In kiwifruit (*A. chinensis*), genes involved in the ascorbic acid biosynthesis and recycling pathway are expanded, which

seems to have resulted from at least one of the two recent WGDs in the kiwifruit genome (53). The same trend was observed in sweet orange (*Citrus sinensis*), in which many of the genes involved in the four known biosynthesis branch pathways of ascorbic acid (and especially in the galacturonate pathway) were upregulated specifically in the fruit (120). The sweet orange genome contains 18 paralogous genes coding for D-galacturonic acid reductase, more than are present in the genomes of apple and grapevine (17 genes); strawberry (15 genes); papaya (13 genes); rice, maize, and cocoa (12 genes); *Arabidopsis* (7 genes); *Brachypodium* (6 genes); and *Chlamydomonas* (no genes). Because these 18 genes were present in two clusters, they probably originated from tandem duplication (120).

Jujube fruit contains even more ascorbic acid than orange and kiwifruit. Analysis of expression during fruit development showed that the more activated genes in jujube ascorbic acid biosynthesis were specific to the L-galactose pathway (71). The gene coding for monodehydroascorbate reductase (the key enzyme in the ascorbic acid recycling pathway) was also overexpressed and significantly expanded compared with other Rosales genomes. Based on phylogenetic analysis of jujube with six other Rosales species (*F. vesca*, *M.* × *domestica*, *M. notabilis*, *P. mume*, *P. persica*, and *P.* × *bretschneideri*) and *Citrus sinensis*, Liu et al. (71) discovered five major monodehydroascorbate reductase gene subfamilies, of which subfamily V is specific to jujube and subfamily IV is specific to Rosaceae. All eight copies of monodehydroascorbate reductase genes of subfamily V are located in two clusters on the jujube genome; therefore, they likely originated from tandem duplication.

Another compound that is presumably beneficial for human health because of its antioxidant effect is resveratrol, which is found in grape skin and consequently in wine. In the grapevine genome, both Jaillon et al. (56) and Velasco et al. (110) observed an expansion of the gene family that codes for stilbene synthases, the resveratrol precursors, although they reported different numbers of predicted genes. Cocoa is one of the richest sources of catechin and epicatechin, two flavonoids with antioxidant activity, and the cacao genome has 18 genes coding for dihydroflavonol-4-reductase (compared with 1 in *Arabidopsis*), an enzyme with a key role in the production of those components (6).

CONCLUSIONS

Access to the genomes of trees is key to the discovery of genes responsible for important agronomic characters, such as fruit quality and biotic- and abiotic-stress resistance, and to the implementation of marker-assisted breeding. As more species are sequenced, comparative analysis among the different genomes will increase and will likely further elucidate the complex basis of traits of interest. These studies will also likely enable the identification of genomic regions linked to the tree-specific phenotypes summarized in this review, whose extended and complex genetic controls are not completely understood.

In recent years, easier access to sequencing technologies and the publication of reference genome sequences have accelerated the development of single-nucleotide-polymorphism arrays, population genomics, and genome-wide association studies. These studies are shedding light on the domestication process of commercially important crops. In fruit trees, the focus is on the development of bigger and more flavored fruits, whereas in forest trees, the aim is to increase the production of wood products and biofuels. As a consequence, genomic regions and candidate genes associated with fruit development and quality, perennial traits, and responses to biotic and abiotic stress have been identified. Continuous improvement of sequencing technologies will allow the publication of higher-quality reference genome sequences, with the same trend that we have observed in the last decade, facilitating analyses of repetitive portions of the genomes and the causes and consequences of different genome sizes.

SUMMARY POINTS

1. The advent of the next-generation-sequencing era has accelerated progress in the sequencing of tree genomes.
2. Both genome size and chromosome number are less variable in trees than they are in herbaceous plants. Genome size is correlated neither with gene content nor with chromosome number in tree species.
3. Whole-genome duplications have been common in angiosperm trees and herbaceous plants but rare in gymnosperms.
4. Although trees differ from nontree plants in their major reproductive strategy, the function of flowering-time genes such as the *CO/FT* regulatory module seems to be conserved between annual plants and trees.
5. Recent expansions and subsequent subfunctionalization of gene families involved in biotic and abiotic stress may have contributed to adaptation to the environment in genera such as *Eucalyptus* and *Populus*.
6. Genome-wide studies of the CBF/DREB cold-stress gene family have shown a different and more complex regulation and impact on abiotic-stress resistance in some trees than in herbaceous plants.
7. Reference tree genome sequences and subsequent studies have revealed that trees have extensive NBS-LRR disease resistance gene families that are considerably larger than those of herbaceous plants.
8. Gene families involved in the synthesis and transport of sugars are often expanded in fruit trees, such as apple, banana, date, jujube, papaya, peach, and pear trees.
9. The jujube, kiwi, and orange fruits contain high levels of ascorbic acid (vitamin C) owing to recent gene expansions in the ascorbic acid pathway.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We wish to acknowledge the US Department of Agriculture's National Institute of Food and Agriculture for funding the research project PineRefSeq (2011-67009-30030) and the support of the California Walnut Board, the California Pear Advisory Board, the Pear Pest Management Research Fund, and the Washington Tree Fruit Research Commission. X.-X.W. is supported by the visiting scholar program of the China Scholarship Council and the National Natural Science Foundation of China (grant numbers 31270422 and 31470316).

LITERATURE CITED

1. Airoldi CA, Davies B. 2012. Gene duplication and the evolution of plant MADS-box transcription factors. *J. Genet. Genom.* 39:157–65

2. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, et al. 2011. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.* 29:521–27
3. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, et al. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* 4:2274
4. Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181:1–20
5. Arabidopsis Genome Initat. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
6. Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, et al. 2011. The genome of *Theobroma cacao*. *Nat. Genet.* 43:101–8
7. Avia K, Kärkkäinen K, Lagercrantz U, Savolainen O. 2014. Association of *FLOWERING LOCUS T/TERMINAL FLOWER 1*-like gene *FTL2* expression with growth rhythm in Scots pine (*Pinus sylvestris*). *New Phytol.* 204:159–70
8. Barghini E, Natali L, Cossu RM, Giordani T, Pindo M, et al. 2014. The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. *Genome Biol. Evol.* 6:776–91
9. Bennett MD, Leitch IJ. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann. Bot.* 107:467–590
10. Bennetzen JL. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115:29–36
11. Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9:1509–14
12. Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95:127–32
13. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* 29:1492–97
14. Böhlenius H, Huang T, Charbonnel-Campaa L, Brunner AM, Jansson S, et al. 2006. *CO/FT* regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* 312:1040–43
15. Cao PB, Azar S, SanClemente H, Mounet F, Dunand C, et al. 2015. Genome-wide analysis of the AP2/ERF family in *Eucalyptus grandis*: an intriguing over-representation of stress-responsive *DREB1/CBF* genes. *PLOS ONE* 10:e0121041
16. Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat. Rev. Genet.* 14:100–12
17. Chagné D. 2015. Whole genome sequencing of fruit tree species. See Ref. 92, pp. 1–37
18. Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, et al. 2014. The draft genome sequence of European pear (*Pyrus communis* L. "Bartlett"). *PLOS ONE* 9:e92644
19. Chardon F, Damerval C. 2005. Phylogenomic analysis of the PEBP gene family in cereals. *J. Mol. Evol.* 61:579–90
20. Chen C-H, Kuo TC-Y, Yang M-H, Chien T-Y, Chu M-J, et al. 2014. Identification of cucurbitacins and assembly of a draft genome for *Aquilaria agallocha*. *BMC Genom.* 15:578
21. Chen J, Tian Q, Pang T, Jiang L, Wu R, et al. 2014. Deep-sequencing transcriptome analysis of low temperature perception in a desert tree, *Populus euphratica*. *BMC Genom.* 15:326
22. Chen SC, Cannon CH, Kua CS, Liu JJ, Galbraith DW. 2014. Genome size variation in the Fagaceae and its implications for trees. *Tree Genet. Genomes* 10:977–88
23. Christie N, Tobias PA, Naidoo S, Külheim C. 2016. The *Eucalyptus grandis NBS-LRR* gene family: physical clustering and expression hotspots. *Front. Plant Sci.* 6:1238
24. Cossu RM, Giordani T, Cavallini A, Natali L. 2014. High-throughput analysis of transcriptome variation during water deficit in a poplar hybrid: a general overview. *Tree Genet. Genomes* 10:53–66
25. D'Hont A, Denoeud F, Aury J, Baurens F, Carreel F, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488:213–19
26. Dai X, Hu Q, Cai Q, Feng K, Ye N, et al. 2014. The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Res.* 24:1274–77

27. Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J. 2013. A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids. *BMC Genom*. 14:683
28. De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, et al. 2014. Insights into conifer giga-genomes. *Plant Physiol*. 166:1–9
29. De La Torre AR, Lin YC, Van De Peer Y, Ingvarsson PK. 2015. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *Picea* gene families. *Genome Biol. Evol.* 7:1002–15
30. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345:1181–84
31. Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis. Genome Res.* 12:1075–79
32. DeYoung BJ, Innes RW. 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* 7:1243–49
33. Ensminger I, Yao-Yun Chang C, Bräutigam K. 2015. Tree responses to environmental cues. See Ref. 92, pp. 229–63
34. Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, et al. 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat. Genet.* 46:1089–96
35. Flavell RB, Bennett MD, Smith JB, Smith DB. 1974. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.* 12:257–69
36. Friedman J, Rubin MJ. 2015. All in good time: understanding annual and perennial strategies in plants. *Am. J. Bot.* 102:497–99
37. Geraldes A, Farzaneh N, Grassa CJ, McKown AD, Guy RD, et al. 2014. Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution* 68:3260–80
38. Giovannoni JJ. 2004. Genetic regulation of fruit development and ripening. *Plant Cell* 16:S170–80
39. Giovannoni JJ. 2007. Fruit ripening mutants yield insights into ripening control. *Curr. Opin. Plant Biol.* 10:283–89
40. Gonzalez-Ibeas D, Martinez-Garcia PJ, Famula RA, Delfino-Mix A, Stevens KA, et al. 2016. Assessing the gene content of the megagenome: sugar pine (*Pinus lambertiana*). *G3* 6:3787–802
41. Gramzow L, Theißen G. 2010. A hitchhiker's guide to the MADS world of plants. *Genome Biol.* 11:214
42. Gramzow L, Theißen G. 2013. Phylogenomics of MADS-box genes in plants—two opposing life styles in one gene family. *Biology* 2:1150–64
43. Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, et al. 2012. Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. *Tree Genet. Genomes* 8:463–508
44. Groover AT. 2005. What genes make a tree a tree? *Trends Plant Sci.* 10:210–14
45. Gyllenstrand N, Clapham D, Källman T, Lagercrantz U, Ka T. 2007. A Norway spruce *FLOWERING LOCUS T* homolog is implicated in control of growth rhythm in conifers. *Plant Physiol.* 144:248–57
46. Harfouche A, Meilan R, Altman A. 2014. Molecular and physiological responses to abiotic stress in forest trees and their relevance to tree improvement. *Tree Physiol.* 34:1181–98
47. He N, Zhang C, Qi X, Zhao S, Tao Y, et al. 2013. Draft genome sequence of the mulberry tree *Morus notabilis. Nat. Commun.* 4:2445
48. Hedman H, Källman T, Lagercrantz U. 2009. Early evolution of the MFT-like gene family in plants. *Plant Mol. Biol.* 70:359–69
49. Hirakawa H, Nakamura Y, Kaneko T, Isobe S, Sakai H, et al. 2011. Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. *Plant Biotechnol.* 28:471–80
50. Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW. 2016. Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *Populus trichocarpa*. *New Phytol.* 209:1240–51
51. Hsu C-Y, Adams JP, Kim H, No K, Ma C, et al. 2011. *FLOWERING LOCUS T* duplication coordinates reproductive and vegetative growth in perennial poplar. *PNAS* 108:10756–61

34. Describes the first large-scale whole-genome resequencing study, which provided insights into the genomics of local adaptation in poplar.

52. Hsu C-Y, Liu Y, Luthe DS, Yuceer C. 2006. Poplar *FT2* shortens the juvenile phase and promotes seasonal flowering. *Plant Cell* 18:1846–61
53. Huang S, Ding J, Deng D, Tang W, Sun H, et al. 2013. Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* 4:2640
54. Hudson CJ, Kullan ARK, Freeman JS, Faria DA, Grattapaglia D, et al. 2012. High synteny and colinearity among eucalyptus genomes revealed by high-density comparative genetic mapping. *Tree Genet. Genomes* 8:339–52
55. Ireland HS, Yao JL, Tomes S, Sutherland PW, Nieuwenhuizen N, et al. 2013. Apple *SEPALLATA1/2*-like genes control fruit flesh development and ripening. *Plant J.* 73:1044–56
56. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–67
57. Karlgren A, Gyllenstrand N, Clapham D, Lagercrantz U. 2013. *FLOWERING LOCUS T/TERMINAL FLOWER1*-like genes affect growth rhythm and bud set in Norway spruce. *Plant Physiol.* 163:792–803
58. Karlgren A, Gyllenstrand N, Källman T, Sundström JF, Moore D, et al. 2011. Evolution of the PEBP gene family in plants: functional diversification in seed plant evolution. *Plant Physiol.* 156:1967–77
59. Karlova R, Chapman N, David K, Angenent GC, Seymour GB, De Maagd RA. 2014. Transcriptional control of fleshy fruit development and ripening. *J. Exp. Bot.* 65:4527–41
60. Kersting AR, Mizrachi E, Bornberg-Bauer E, Myburg AA. 2015. Protein domain evolution is associated with reproductive diversification and adaptive radiation in the genus *Eucalyptus*. *New Phytol.* 206:1328–36
61. Klee HJ. 2010. Improving the flavor of fresh fruits: genomics, biochemistry, and biotechnology. *New Phytol.* 187:44–56
62. Klintenäs M, Pin PA, Benlloch R, Ingvarsson PK, Nilsson O. 2012. Analysis of conifer *FLOWERING LOCUS T/TERMINAL FLOWER1*-like genes provides evidence for dramatic biochemical evolution in the angiosperm *FT* lineage. *New Phytol.* 196:1260–73
63. Knapp S, Litt A. 2013. Fruit—an angiosperm innovation. In *The Molecular Biology and Biochemistry of Fruit Ripening*, ed. GB Seymour, M Poole, JJ Giovannoni, GA Tucker, pp. 21–42. Ames, IA: Wiley-Blackwell
64. Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, et al. 2008. Genome-wide identification of *NBS* resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* 66:619–36
65. Krishnan NM, Pattnaik S, Jain P, Gaur P, Choudhary R, et al. 2012. A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm *Azadirachta indica*. *BMC Genom.* 13:464
66. Lau N, Makita Y, Kawashima M, Taylor TD, Kondo S, et al. 2016. The rubber tree genome shows expansion of gene family associated with rubber biosynthesis. *Sci. Rep.* 6:28594
67. Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* 320:481–83
68. Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytol.* 194:629–46
69. Lesur I, Le Provost G, Bento P, Da Silva C, Leplé J-C, et al. 2015. The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. *BMC Genom.* 16:112
70. Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, et al. 2015. Early genome duplications in conifers and other seed plants. *Sci. Adv.* 1:e1501084
71. Liu M-J, Zhao J, Cai Q-L, Liu G-C, Wang J-R, et al. 2014. The complex jujube genome provides insights into fruit tree biology. *Nat. Commun.* 5:5315
72. Liu Y-Y, Yang K-Z, Wei X-X, Wang X-Q. 2016. Revisiting the phosphatidylethanolamine-binding protein (PEBP) gene family reveals cryptic *FLOWERING LOCUS T* gene homologs in gymnosperms and sheds new light on functional evolution. *New Phytol.* 212:730–44
73. Loescher WH. 1987. Physiology and metabolism of sugar alcohols in higher plants. *Physiol. Plant.* 70:553–57
74. Ma T, Wang J, Zhou G, Yue Z, Hu Q, et al. 2013. Genomic insights into salt adaptation in a desert poplar. *Nat. Commun.* 4:2797

56. Describes the sequenced genome with the most conserved arrangement from the paleohexaploid from which all eudicots originated.

75. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, et al. 2016. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of nonstructural polyphenols. *Plant J.* 87:507–32
76. Matzke MA, Mosher RA. 2014. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* 15:394–408
77. McAtee PA, Richardson AC, Nieuwenhuizen NJ, Gunaseelan K, Hoong L, et al. 2015. The hybrid non-ethylene and ethylene ripening response in kiwifruit (*Actinidia chinensis*) is associated with differential regulation of mads-box transcription factors. *BMC Plant Biol.* 15:304
78. McCourt RM, Chapman RL, Buchheim M, Mishler BD. 1996. Green plants. *Tree of Life Web Project*. **http://www.tolweb.org/Green_plants**
79. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991–96
80. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, et al. 2009. Evolution of genome size and complexity in *Pinus*. *PLOS ONE* 4:e4332
81. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, et al. 2014. The genome of *Eucalyptus grandis*. *Nature* 510:356
82. Navarro M, Marque G, Ayax C, Keller G, Borges JP, et al. 2009. Complementary regulation of four *Eucalyptus CBF* genes under various cold conditions. *J. Exp. Bot.* 60:1–12
83. Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* 12:111–22
84. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* 15:R59
85. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–84
86. Palazzo AF, Gregory TR. 2014. The case for junk DNA. *PLOS Genet.* 10:e1004351
87. Parent GJ, Raherison E, Sena J, Mackay JJ. 2015. Forest tree genomics: review of progress. See Ref. 92, pp. 39–92
88. Park WA, Allen CD, Macalady AK, Griffin D, Woodhouse CA, et al. 2013. Temperature as a potent driver of regional forest drought stress and tree mortality. *Nat. Clim. Change* 3:292–97
89. Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. 2012. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol.* 10:84
90. Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* 164:10–15
91. Perazzolli M, Malacarne G, Baldo A, Righetti L, Bailey A, et al. 2014. Characterization of resistance gene analogues (*RGA*s) in apple (*Malus* × *domestica* Borkh.) and their evolutionary history of the Rosaceae family. *PLOS ONE* 9:e83844
92. Plomion C, Adam-Blondonpp A-F, eds. 2015. *Land Plants - Trees*. Adv. Bot. Res. Vol. 74. London: Academic
93. Plomion C, Aury JM, Amselem J, Alaeitabar T, Barbe V, et al. 2016. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Mol. Ecol. Resour.* 16:254–65
94. Plomion C, Bastien C, Bogeat-Triboulot M-B, Bouffier L, Déjardin A, et al. 2016. Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Ann. For. Sci.* 73:77–103
95. Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, et al. 2007. Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* 266:5–43
96. Quinn JJ, Chang HY. 2015. Unique features of long non-coding rna biogenesis and function. *Nat. Rev. Genet.* 17:47–62
97. Rahman AYA, Usharraj AO, Misra BB, Thottathil GP, Jayasejaran K, et al. 2013. Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genom.* 14:75
98. Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, et al. 2011. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res.* 18:65–76
99. Scott AD, Stenz NWM, Ingvarsson PK, Baum DA. 2016. Whole genome duplication in coast redwood (*Sequoia sempervirens*) and its implications for explaining the rarity of polyploidy in conifers. *New Phytol.* 211:186–93

81. Describes a comprehensive study of gene families involved in lignocellulosic biomass production and secondary metabolism and oils in *Eucalyptus*

84. Describes the first *Pinus* species sequenced and provides insights into the genomic organization of gymnosperms and how they differ from flowering plants.

85. Provides the first comparative genomics study of transposable elements in gymnosperm species.

100. Seymour GB, Ostergaard L, Chapman NH, Knapp S, Martin C. 2013. Fruit development and ripening. *Annu. Rev. Plant Biol.* 64:219–41
101. Shuai P, Liang D, Tang S, Zhang Z, Ye CY, et al. 2014. Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in *Populus trichocarpa*. *J. Exp. Bot.* 65:4975–83
102. Singh R, Ong-Abdullah M, Low E-TL, Manaf MAA, Rosli R, et al. 2013. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature* 500:335–39
103. Soltis DE, Soltis PS, Bennett MD, Leitch IJ. 2003. Evolution of genome size in the angiosperms. *Am. J. Bot.* 90:1596–603
104. Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, et al. 2016. Sequence of the sugar pine megagenome. *Genetics* 204:1613–26
105. Tang C, Yang M, Fang Y, Luo Y, Gao S, et al. 2016. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat. Plants* 2:1–10
106. Tang S, Dong Y, Liang D, Zhang Z, Ye CY, et al. 2015. Analysis of the drought stress-responsive transcriptome of black cottonwood (*Populus trichocarpa*) using deep RNA sequencing. *Plant Mol. Biol. Rep.* 33:424–38
107. Tang S, Liang H, Yan D, Zhao Y, Han X, et al. 2013. *Populus euphratica*: the transcriptomic response to drought stress. *Plant Mol. Biol.* 83:539–57
108. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–604
109. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* 42:833–39
110. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLOS ONE* 2:e1326
111. Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, et al. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45:487–94
112. Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, et al. 2002. A MADS-box gene necessary for fruit ripening at the tomato *ripening-inhibitor* (*rin*) locus. *Science* 296:343–46
113. Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, et al. 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species rad markers. *Mol. Ecol.* 22:3098–111
114. Wang Y, Hu Z, Yang Y, Chen X, Chen G. 2010. Genome-wide identification, phylogeny, and expression analysis of the SBP-box gene family in grapevine. *Russ. J. Plant Physiol.* 57:273–82
115. Warren RL, Keeling CI, Saint Yuen MM, Raymond A, Taylor GA, et al. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* 83:189–212
116. Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, et al. 2014. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* 196:891–909
117. Wisniewski M, Nassuth A, Teulières C, Marque C, Rowland J, et al. 2014. Genomics of cold hardiness in woody plants. *CRC Crit. Rev. Plant Sci.* 33:92–124
118. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, et al. 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* 32:656–62
119. Wu J, Wang Z, Shi Z, Zhang S, Ming R, et al. 2013. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* 23:396–408
120. Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, et al. 2013. The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* 45:59–66
121. Yamamoto T, Kimura T, Saito T, Kotobuki K, Matsuta N, et al. 2004. Genetic linkage maps of Japanese and European pears aligned to the apple consensus map. *Acta Hortic* 663:51–56
122. Yang S, Feng Z, Zhang X, Jiang K, Jin X, et al. 2006. Genome-wide investigation on the genetic variations of rice disease resistance genes. *Plant Mol. Biol.* 62:181–93
123. Yang S, Zhang X, Yue JX, Tian D, Chen JQ. 2008. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genom.* 280:187–98

108. Describes the first tree genome sequenced and the first comparative genomics study of a tree species (poplar) and a nontree species (*Arabidopsis*).

109. Provides evidence that the expansion of MADS-box genes in the apple genome may be related to the development of the pome.

111. Describes a high-quality assembly of a double haploid that has been used as a reference for studies in several other *Prunus* species.

124. Yoon S-K, Park E-J, Choi Y-I, Bae E-K, Kim J-H, et al. 2014. Response to drought and salt stress in leaves of poplar (*Populus alba* × *Populus glandulosa*): expression profiling by oligonucleotide microarray analysis. *Plant Physiol. Biochem.* 84:158–68
125. Zhang J, Jiang D, Liu B, Luo W, Lu J, et al. 2014. Transcriptome dynamics of a desert poplar (*Populus pruinosa*) in response to continuous salinity stress. *Plant Cell Rep.* 33:1565–79
126. Zhang Q, Chen W, Sun L, Zhao F, Huang B, et al. 2012. The genome of *Prunus mume*. *Nat. Commun.* 3:1318
127. Zhao T, Liang D, Wang P, Liu J, Ma F. 2012. Genome-wide analysis and expression profiling of the dreb transcription factor gene family in *Malus* under abiotic stress. *Mol. Genet. Genom.* 287:423–36
128. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, et al. 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* 196:875–90

Annual Review of Plant Biology
Volume 68, 2017

Contents

Errata

An online log of corrections to *Annual Review of Plant Biology* articles may be found at http://www.annualreviews.org/errata/arplant