Step-wise and lineage-specific diversification of plant RNA polymerase genes and origin of the largest plant-specific subunits

Yaqiong Wang[1,2] **and Hong Ma**[1,2,3]

[1]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Institute of Plant Biology, Center for Evolutionary Biology, School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China; [2]Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering and Institute of Biodiversity Sciences, Fudan University, Shanghai 200433, China; [3]Institutes of Biomedical Sciences, Fudan University, 138 Yixueyuan Road, Shanghai 200032, China

Author for correspondence:
Hong Ma
Tel: +86 21 51630532
Email: hongma@fudan.edu.cn

Received: *8 February 2015*
Accepted: *24 March 2015*

New Phytologist (2015) **207:** 1198–1212
doi: 10.1111/nph.13432

Key words: gene duplication, parallel evolution, Pol IV, Pol V, protein complex, RNA polymerase.

Summary

- Proteins often function as complexes, yet little is known about the evolution of dissimilar subunits of complexes. DNA-directed RNA polymerases (RNAPs) are multisubunit complexes, with distinct eukaryotic types for different classes of transcripts. In addition to Pol I–III, common in eukaryotes, plants have Pol IV and V for epigenetic regulation. Some RNAP subunits are specific to one type, whereas other subunits are shared by multiple types.
- We have conducted extensive phylogenetic and sequence analyses, and have placed RNAP gene duplication events in land plant history, thereby reconstructing the subunit compositions of the novel RNAPs during land plant evolution.
- We found that Pol IV/V have experienced step-wise duplication and diversification of various subunits, with increasingly distinctive subunit compositions. Also, lineage-specific duplications have further increased RNAP complexity with distinct copies in different plant families and varying divergence for subunits of different RNAPs. Further, the largest subunits of Pol IV/V probably originated from a gene fusion in the ancestral land plants.
- We propose a framework of plant RNAP evolution, providing an excellent model for protein complex evolution.

Introduction

Most proteins accomplish their functions through interaction networks or as subunits in protein complexes. The BioGRID (3.2.120) database records 6439, 7184 and 19 746 proteins that interact physically with other proteins as supported experimentally in yeast, *Arabidopsis* and humans, respectively (http://wiki.thebiogrid.org/doku.php/statistics). Protein complexes have many essential cellular functions, such as DNA polymerases in replication (Kelman & O'Donnell, 1995), RNA polymerases (RNAPs) and the mediator complex in transcription and its regulation (Kelleher *et al.*, 1990; Cramer *et al.*, 2001), ribosome and proteasome in protein synthesis and degradation (Ben-Shem *et al.*, 2011; Beck *et al.*, 2012), ATP synthase in energy metabolism (Boyer, 1997) and the G protein complex in signaling (Neer, 1995), just to name a few.

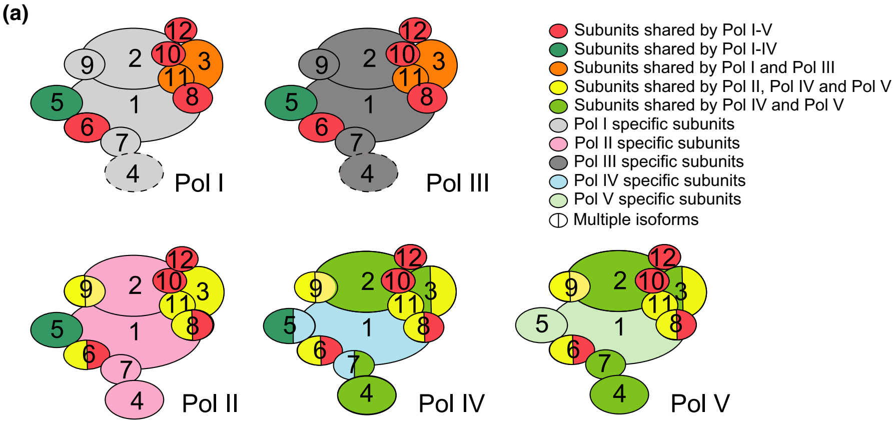
Gene duplication is a major mechanism for increasing network complexity (Force *et al.*, 1999; Innan & Kondrashov, 2010; Baker *et al.*, 2013). Although many duplicates (paralogs) are lost after duplications, some undergo subfunctionalization, with partial retention of ancestral functions, whereas others are maintained after neofunctionalization (acquisition of new functions) (Lynch & Conery, 2000; Moore & Purugganan, 2005). Duplication patterns in individual gene families have been investigated extensively (Nei & Rooney, 2005; Lin *et al.*, 2006, 2007; Nei *et al.*, 2008; Zhou & Ma, 2008; Xu *et al.*, 2009), but studies of genes encoding nonhomologous subunits of complexes are very limited. For example, it is unclear whether genes encoding subunits of a complex are duplicated and retained at nearly the same time to achieve functional diversification. Alternatively, genes for different subunits might duplicate at different evolutionary times, highlighting step-wise or progressive functional evolution. Therefore, it is important to examine the evolutionary histories of members of the same complex to gain insights into similarities and differences between subunits of complexes and their potential functional impact.

DNA-directed RNAPs are complex molecular machines for the essential function of the synthesis of RNA from DNA templates (Fig. 1a). All eukaryotes examined so far have three types of nuclear RNAP: Pol I, II and III for rRNA, mRNAs and noncoding RNAs, and tRNAs and 5S RNAs, respectively (Roeder & Rutter, 1969; Vannini & Cramer, 2012). Recently, two additional RNAPs (Pol IV and Pol V) have been reported in plants (*Arabidopsis* and maize); they are important for epigenetic regulations, such as small interfering RNA (siRNA)-directed DNA methylation (RdDM) and gene silencing (Onodera *et al.*, 2005;

Erhard *et al.*, 2009; Haag & Pikaard, 2011). In *Arabidopsis*, Pol IV is required for siRNA biogenesis; Pol V transcribes scaffold RNA, thereby recruiting the silencing complex (Haag & Pikaard, 2011; Pikaard *et al.*, 2012). In maize, Pol IV is also important for the regulation of development and paramutations (Erhard *et al.*, 2009, 2013; Pikaard & Tucker, 2009; Stonaker *et al.*, 2009).

Eukaryotic RNAPs consist of 12–17 subunits of varying sequences and sizes (12 of the subunits are illustrated in Fig. 1a). Structural studies have revealed that the *Saccharomyces cerevisiae* (yeast) Pol I, II and III have similar structures, with a 10-subunit core and two to seven peripheral components (Cramer *et al.*, 2008). By convention, the subunits are named using 'RP' for RNAP, 'A', 'B' and 'C' for Pol I, II and III, respectively, and a number (starting from the largest) for the specific subunit; for example, RPA1 is the 1st (largest) subunit for Pol I. Most yeast Pol II subunits are essential for cell viability, except for the 4th (RPB4) and 9th (RPB9) subunits (Archambault & Friesen, 1993; Hull *et al.*, 1995; Sampath & Sadhale, 2005). Within the conserved core, five subunits (1st, 2nd, 3rd, 9th, 11th) are each encoded by a multigene family, having specific paralogs for Pol I, II and III, respectively (Cramer *et al.*,

(a)


12
2
10
3
9
11
5
1
8
6
7
4
Pol I
Pol III
Pol II
Pol IV
Pol V
Subunits shared by Pol I-V
Subunits shared by Pol I-IV
Subunits shared by Pol I and Pol III
Subunits shared by Pol II, Pol IV and Pol V
Subunits shared by Pol IV and Pol V
Pol I specific subunits
Pol II specific subunits
Pol III specific subunits
Pol IV specific subunits
Pol V specific subunits
Multiple isoforms

(b)

	Type	a					β		γ				
	Subunit	1	2	4	7	9	3	11	5	6	8	10	12
Plants	*Arabidopsis thaliana* (Thale cress)	5	5	2*(?)	6	6	4	2	6	2	2	2	2
	Mimulus guttatus (Monkey flower)	6	6	2*(?)	9	3	2	4	4	1	2	1	2
	Zea mays (Maize)	6	7	2*(?)	4	5	2	6	3	2	2	2	3
	Amborella trichopoda	5	4	3*(?)	5	3	2	2	3	1	1	2	2
	Selaginella moellendorffii (Spikemoss)	5	6	1*(?)	5	6	2	3	1	2	2	1	2
	Physcomitrella patens (Moss)	7	5	1*(?)	4	3	3	3	3	1	1	1	1
	Chlamydomonas reinhardtii (Green algae)	3	3	1*(?)	3	3	2	2	1	1	1	1	1
Animals	*Homo sapiens* (Human)	3	3	2†(?)	3	3	2	2	1	1	1	1	1
	Drosophila melanogaster (Fruit fly)	3	3	2†(?)	3	3	2	2	1	1	1	1	1
Fungi	*Neurospora crassa*	3	3	3	3	3	2	2	1	1	1	1	1
	Saccharomyces cerevisiae *(Budding yeast)*	3	3	3	3	3	2	2	1	1	1	1	1

Fig. 1 Features of RNA polymerase (RNAP) subunits in *Arabidopsis* and other representative species. (a) Subunit compositions of RNAPs in *Arabidopsis*. The compositions of Pol II, IV and V subunits were from previous biochemical studies (Huang *et al.*, 2009; Ream *et al.*, 2009; Law *et al.*, 2011). Pol I and Pol III subunits are classified according to gene annotations of the *Arabidopsis* genome. The 4th subunits of Pol I and III are shown with dashed outlines to indicate the lack of recognized *Arabidopsis* genes. (b) Copy numbers of eukaryotic RNAP genes. A darker color indicates a higher copy number. †, Only numbers for the Pol II and Pol III 4th subunits are shown. For animals, the putative counterparts of Pol I 4th subunits are indicated by a question mark. The yeast Pol I, II and III 4th subunits are RPA14, RPB4 and RPC17, respectively (Cramer *et al.*, 2008). RPC17 shares with RPB4 an RNA_pol_Rpb4 domain (PF03874 in the Pfam database) and has a human homolog (CRCP; ENSG00000241258). However, RPA14 has no detectable amino acid sequence similarity to RPB4 or detectable homologs in animals. *, Only numbers for the Pol II and Pol IV/V 4th subunits are shown. For plants, the putative homologs of Pol I and Pol III 4th subunits are indicated by a question mark. *Arabidopsis* has two putative homologs of RPC17 with *c.* 53% similarity for *c.* 40% of the amino acid sequences with high E values (0.003 and 0.007; much higher than the 10^{-5} threshold), but no detectable homologs of RPA14.

2008), and probably with distinct functions for each type of RNAP, such as template recognition. Each of five other subunits is shared by all three types of RNAP and encoded by a single gene, with names RPB5, RPB6, RPB8, RPB10 and RPB12, respectively (as for Pol II) (Cramer *et al.*, 2008; Vannini & Cramer, 2012), probably with the same functions for all three RNAP types.

Unlike animals and fungi, plants have multiple genes for nearly all subunits (Luo & Hall, 2007; Tucker *et al.*, 2011). In *Arabidopsis*, the names of RNAP subunits start with an 'N' for nuclear RNAPs; it has been established biochemically that Pol II, IV and V each have a similar 12-subunit structure and share the 3rd, 6th, 8th, 9th, 10th, 11th and 12th subunits (Ream *et al.*, 2009) (Supporting Information Table S1). Pol IV and V have several subunits distinct from the Pol II counterparts: different Pol IV and Pol V 1st subunits; the 2nd (NRPD2), 4th (NRPD4) and 7th (NRPE7) subunits are shared by Pol IV and V, but different from Pol II, and an additional 7th subunit (NRPD7) for Pol IV; Pol II and IV share one copy of the 5th subunit, but both Pol IV and Pol V have their specific 5th subunit, respectively; and in addition to the shared 3rd subunit (NRPB3) between Pol II, IV and V, Pol IV and Pol V also have another copy (NRPE3B). *Arabidopsis* mutants defective for one of several Pol I, II and III subunits (e.g. *nrpa2*, *nrpb2*, *nrpc2*) showed female gametophyte lethality (Onodera *et al.*, 2008). By contrast, mutants defective in the largest subunits of Pol IV and V (*NRPD1* and *NRPE1*) were viable, but flowered late under short-day conditions (Onodera *et al.*, 2005; Lahmy *et al.*, 2009; Ream *et al.*, 2009).

Previously, analyses of genes from a small number of plant species have suggested that the common ancestral gene for the Pol IV/V 1st subunit probably originated before the divergence of land plants (Luo & Hall, 2007; Tucker *et al.*, 2011), but when the Pol IV and V 1st subunits diverged is unclear. The genes for the 4th and 5th subunits of Pol IV/V experienced duplication after seed plants diverged from moss (Tucker *et al.*, 2011), but whether the duplication was before or after the divergence of angiosperms and gymnosperms is not clear. In addition, some plant RNAP subunit genes show additional duplications resulting in new combination types (Luo & Hall, 2007; Lahmy *et al.*, 2009; Pikaard & Tucker, 2009; Ream *et al.*, 2009; Tucker *et al.*, 2011). However, there has been no systematic analysis of the evolution of all 12 subunits for Pol II, IV and V during land plant history.

To systematically investigate the origins, duplication and loss patterns, and sequence divergence of RNAP genes in eukaryotes and, particularly, in land plants, we obtained 2228 sequences from 58 eukaryotes and performed comprehensive phylogenetic studies of RNAP genes in representative eukaryotes and major lineages of land plants. Our analyses of the plant RNAP genes provide a comprehensive evolutionary portrait of the conservation and divergence of all 12 subunits, indicating that RNAPs progressively acquired different new functions in evolution by having new genes for various subunits in Pol IV and V at different times. In addition, different angiosperm groups experienced lineage-specific duplications for several subunits, suggesting that they functionally diverged independently. We further uncovered that the largest subunits of Pol IV/V probably originated from a gene fusion event. These results suggest that the functions of RNAPs are probably more diverse among plants than previously realized and provide a general model for the evolution of multiprotein complexes.

Materials and Methods

Retrieval of sequences

Arabidopsis thaliana (L.) Heynh. and *Saccharomyces cerevisiae* RNAP genes were identified previously or in The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org) and Saccharomyces Genome Database (SGD) (www.yeastgenome.org) (Table S1). Protein sequence queries were used to search for homologs by BLASTP or TBLASTN with an E value of $< 1 \times 10^{-5}$. Selected plant, animal and fungus sequences were downloaded from JGI Phytozome v10 (Goodstein *et al.*, 2012), Ensembl (release 64) (ftp://ftp.ensembl.org/), fungal genome databases (fungalgenomes.org/data/) and other databases (Table S2).

Gene nomenclature

Each RNAP gene contains a three-letter species designation from the first letter of the genus and the first two letters of the species (Table S2), with an exception of *Schizosaccharomyces pombe* (fission yeast), abbreviated as 'Scp', to distinguish from 'Spo' for *Spirodela polyrhiza* (duckweed). The species designation is followed by the name of the *Arabidopsis* ortholog (or the most similar for lineage-specific duplicates) (Table S3). The names of animal and fungal RNAP genes contain the three-letter species designation and the name of the *Saccharomyces cerevisiae* ortholog. Additional variants of a subunit are indicated by a lowercase suffix (Table S3).

Phylogenetic analyses

Protein sequences of each subunit family were aligned by Muscle 3.7 (Edgar, 2004) and manually adjusted using Jalview 2.8 (Waterhouse *et al.*, 2009). Maximum likelihood (ML) analysis was performed using RAxML 8.0.0 (Stamatakis, 2014) and rapid bootstrap analysis was performed with the bootstrap convergence test using the extended majority-rule consensus tree criterion (autoMRE) in RAxML. Bayesian analysis (BA) was conducted using MrBayes v3.2.2 (Ronquist *et al.*, 2012) with 10^5 generation runs, four Markov chains and sampling every 500 generations. For each phylogenetic analysis, best-fit evolutionary models were selected by Prottest 3 (Darriba *et al.*, 2011) under the Bayesian information criterion (BIC) (Table S4). Phylogenies of Brassicaceae and Poaceae genes for each subunit were built on alignment of CDS (nucleotide) instead of amino acid sequences, because of the limited information of highly similar protein sequences. DNA sequences were aligned by pal2nal v12.2 (Suyama *et al.*, 2006) on the basis of corresponding protein alignment. The GTR + G model was used in Bayesian and ML analyses. The resulting trees were visualized and adjusted by MEGA 6.0

(Tamura *et al.*, 2013) or FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

Sequence comparisons of *Physcomitrella NRPD1/NRPE1* genes

The protein sequences of the three *Physcomitrella patens* NRPD1/NRPE1 homologs were aligned with the angiosperm NRPD1 and NRPE1 sequences, respectively, by BLASTP. Pp1s193_6 has more identical and similar sites to Ath.NRPD1 (380, 608) and Aco.NRPD1 (367, 593) than it does to Ath.NRPE1 (343, 570) and Aco.NRPE1 (361, 568), suggesting that Pp1s193_6 is an NRPD1 ortholog. However, Pp1s83_67 and Pp1s83_168 are more similar to NRPE1 (Pp1s83_67, 345 identical and 592 similar sites; Pp1s83_186, 414 identical and 658 similar sites) than to Ath.NRPD1 (Pp1s83_67, 315 identical and 543 similar sites; Pp1s83_186, 303 identical and 495 similar sites).

Detection of gene duplication and loss events

ML and Bayesian phylogenies of Brassicaceae and Poaceae RNAP genes were reconciled with species trees of Brassicaceae (Tree topology in newick format: (((((Ath, Aly), ((Cru, Cgr), Bst)), (Esa, Bra)), Cpa), Ptr);) and Poaceae ((((((Sbi, Zma), (Sit, Pvi)), (Osa, Bdi)), Mac), Spo);), respectively, by Notung 2.6 (Chen *et al.*, 2000) (see Table S2 for species names). The minimal number of gene duplication and loss events was detected by the 'rearrange' mode in Notung 2.6: well-supported branches (Bayesian posterior probability values of > 0.8 and ML bootstrap values of > 80) were preserved and weak supported branches were rearranged to minimize the number of duplications.

Synteny analysis

All against all BLASTP search was performed for each proteomic dataset of the seven species of Brassicaceae and the six species of Poaceae. MCScanX (Wang *et al.*, 2012) was used to detect syntenic blocks and the duplicate_gene_classifier program in the MCScanX package was employed to detect syntenic genes probably from whole-genome or segmental duplications.

Selection pressure analysis

The ancestral DNA sequences for each subunit in the most recent common ancestor of Brassicaceae or Poaceae were reconstructed by FastML v3.1 (Pupko *et al.*, 2000) with the joint reconstruction method from alignments of Brassicaceae and Poaceae genes and their corresponding ML trees. Then, the Yang & Nielsen (2000) method, implemented in the yn00 program of the PAML package (Yang, 2007), was used to calculate the nonsynonymous to synonymous rate ratio ($\omega = d_N/d_S$) between each gene and its ancestral sequence. The distributions of d_N/d_S values for each subtype of RNAP subunit were plotted by the boxplot function in R (R Core Team, 2014). Extreme values outside the 1.5-fold of the interquartile range were defined as outliers. As comparisons, the same methods were applied to RNAP genes from eight species of Catarrhini (human and closest relatives) and five species of *Saccharomyces* (Table S2).

Expression analysis

Normalized expression data of genes for each RNAP subunit from 11 *Arabidopsis* tissues were obtained from the At-TAX tilling array dataset (Laubinger *et al.*, 2008). Expression data for 60 maize tissues were retrieved from the ZM37 dataset on plexdb (Dash *et al.*, 2012). Expression levels were visualized using the pheatmap package (Kolde, 2013) in R.

Gene structure analysis

Intron positions were retrieved from genomic GFF files and converted to relative coordinates of the open reading frame (ORF). Phase 0 is for an intron between two codons, phase 1 between the first and second nucleotide of a codon and, otherwise, phase 2.

Results

Land plants have more RNAP genes than others

To obtain sequences for 12 subunits of each RNAP, we searched databases using *Arabidopsis* and yeast RNAP genes as queries (see the Materials and Methods section) and identified 89–319 homologs for 12 subunits of all RNAPs from each of 58 eukaryotes, especially animals, plants and fungi (Table S3). The number of genes for each subunit is constant among the animals, fungi and green algae examined here (Fig. 1b): three copies for each of the 1st, 2nd, 4th, 7th and 9th subunits; two copies for each of the 3rd and 11th subunits; and one copy of the 5th, 6th, 8th, 10th and 12th subunits. However, copy numbers increased to different extents in land plants (Fig. 1b): the copy numbers for moss genes for the 1st, 2nd and 7th subunits were seven, five and four, respectively; *Amborella trichopoda* (the sister of all other angiosperms) also showed increased gene copies for the 1st, 2nd, 4th, 5th, 7th, 10th and 12th subunits, suggesting distinct gene evolutionary patterns of different subunit genes.

Phylogenetic analyses of RNAP subunits show three evolutionary patterns

Genes for each of the 12 RNAP subunits form a distinct family. To investigate their eukaryotic evolutionary histories, we constructed separate phylogenies with sequences from nine plants, four animals and five fungi (Figs 2, S1–S6). We grouped the 12 phylogenies into three types on the basis of their inclusion in eukaryotic RNAPs (Fig. 3): type α subunits are different among Pol I, Pol II and Pol III, with at least three copies in the eukaryotic ancestors (1st, 2nd, 4th, 7th and 9th) (Figs 2a, S1–S4); the type β subunit has two copies in the eukaryotic ancestor and one copy shared by Pol I and Pol III (3rd and 11th) (Figs 2b, S5); and type γ has only one copy (5th, 6th, 8th, 10th and 12th) (Figs 2c, S6).

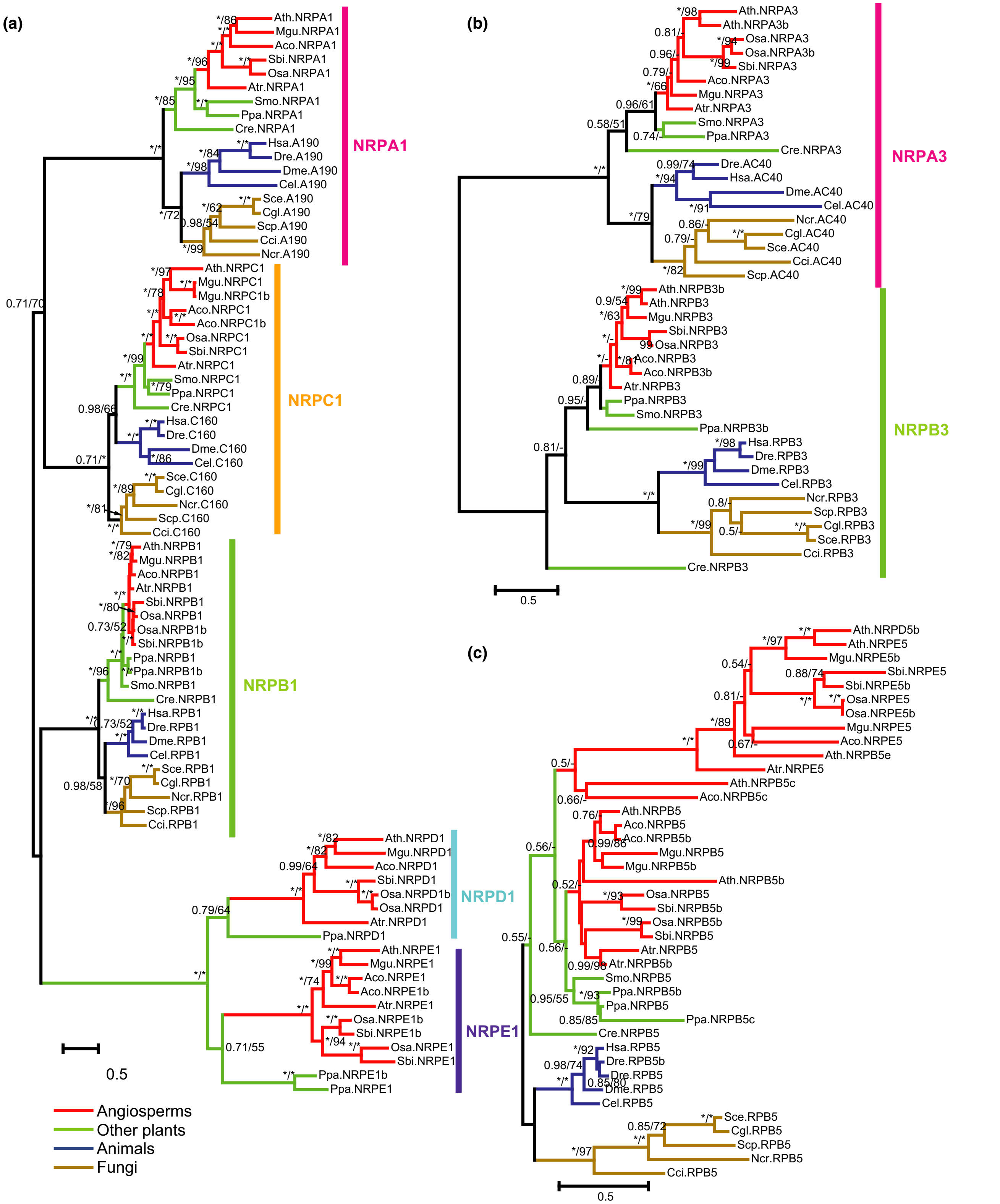


Fig. 2 Representative phylogenies of genes for eukaryotic RNA polymerase (RNAP) subunits for three evolutionary types: (a) 1st subunits (type α); (b) 3rd subunits (type β); (c) 5th subunits (type γ). Tree topologies generated by RAxML are shown here. Bayesian posterior probability values (> 0.5) according to MrBayes and bootstrap values (> 50) from RAxML are labeled on internal nodes. Asterisks (*) indicate Bayesian posterior probability values of 1 or bootstrap values of 100. Aco, *Aquilegia coerulea*; Ath, *Arabidopsis thaliana*; Atr, *Amborella trichopoda*; Cci, *Coprinus cinereus*; Cel, *Caenorhabditis elegans*; Cgl, *Candida glabrata*; Cre, *Chlamydomonas reinhardtii*; Dme, *Drosophila melanogaster*; Dre, *Danio rerio*; Hsa, *Homo sapiens*; Mgu, *Mimulus guttatus*; Ncr, *Neurospora crassa*; Osa, *Oryza sativa*; Ppa, *Physcomitrella patens*; Sbi, *Sorghum bicolor*; Sce, *Saccharomyces cerevisiae*; Scp, *Schizosaccharomyces pombe*; Smo, *Selaginella moellendorffii*.

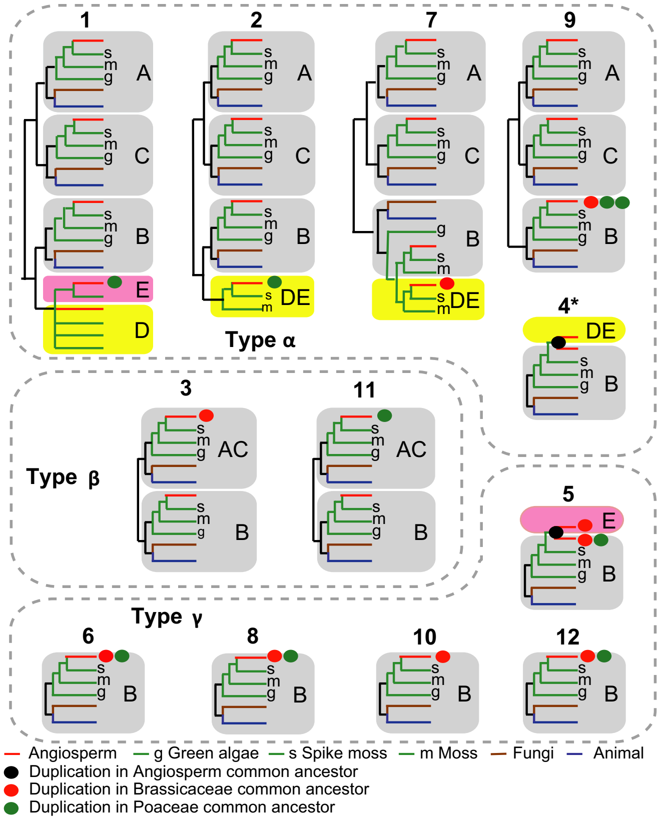

1
2
7
9
s
m
g
A
C
B
E
D
DE
Type α
4*
DE
3
11
AC
B
Type β
5
E
Type γ
6
8
10
12
Angiosperm
g Green algae
s Spike moss
m Moss
Fungi
Animal
Duplication in Angiosperm common ancestor
Duplication in Brassicaceae common ancestor
Duplication in Poaceae common ancestor

Fig. 3 Schematic phylogenetic diagrams of RNA polymerase (RNAP) subunits. The topologies are shown for each subunit from maximum likelihood (ML) and Bayesian (BA) analysis. Type α topologies have three detected copies in the eukaryotic common ancestor. Type β topologies have two detected copies in the eukaryotic common ancestor. Type γ topologies have one copy in the eukaryotic common ancestor. Gray blocks indicate Pol I, II or III clades, whereas yellow (Pol IV or Pol IV/V) and pink (Pol V specific) blocks indicate plant-specific clades. *, Only the Pol II and Pol IV/V 4th subunits are shown.

To further investigate the evolutionary patterns of genes for plant-specific RNAP subunits, we generated gene trees for each of the 1st, 2nd, 5th, 7th and 9th subunits of Pol II, IV and V (Figs 4, S7–S9). To examine more recent evolutionary relationships within various subfamilies of RNAP genes, we constructed phylogenies of genes for each subunit from seven Brassicaceae (Crucifer) species using papaya and poplar as outgroups (Figs S10, S11). As a comparison, we also constructed phylogenetic trees of grass (Poaceae) genes using *Musa acuminata* (banana) and *Spirodela polyrhiza* (duckweed) sequences as outgroups (Figs S10, S11). The detailed results are presented in the following subsections. Lineage-specific duplications in Brassicaceae or Poaceae are referred to as independent duplications.

Type α: genes for the 1st, 2nd, 4th, 7th and 9th subunits Phylogenetic analyses indicate that genes for each of the 1st, 2nd, 7th and 9th subunits form three ancestral groups each shared by animals, plants and fungi: one including genes for Pol I subunits, and another for genes encoding Pol III subunits. The copy numbers for these Pol I and III genes are constant for most of plant history, except for recent independent duplications in some species (Figs 2a, S1–S4, S10, S11). The third group in each family contains genes encoding eukaryotic Pol II subunits, as well as genes for plant-specific Pol IV and V subunits, indicating that Pol IV/V genes are derived from ancestral Pol II genes.

The duplication patterns related to Pol II, IV and V vary among the gene families (Figs 4, S7, S8). The genes for the largest subunit form two highly supported clades (Fig. 2a): one for Pol II and the other for Pol IV and V. The Pol IV/V clade includes genes from moss and other land plants, with angiosperm genes forming two monophyletic groups corresponding to Pol IV and V, respectively, indicating that genes for the largest subunits of Pol IV and Pol V separated in the common ancestor of extant angiosperms (Fig. 4a). Previously, only one gene coding for the Pol IV largest subunit (NRPD1) was obtained from spike moss (*Selaginella*) and moss (*Physcomitrella*) (Luo & Hall, 2007). However, we detected two copies in *Selaginella* and three copies (Pp1s193_6, Pp1s83_67, Pp1s83_186 in version 1.6 annotation) in *Physcomitrella* (Table S3). Their phylogenetic positions are uncertain because of a high degree of sequence divergence; nevertheless, when we excluded *Selaginella* sequences, one moss gene grouped with angiosperm *NRPD1* genes and the other two genes were close to angiosperm *NRPE1* genes (Fig. 4a). In addition, sequence comparison (see the Materials and Methods section) indicates that the *Physcomitrella* protein Pp1s193_6 is most similar to the angiosperm NRPD1 protein, whereas Pp1s83_67 and Pp1s83_168 are most similar to NRPE1, supporting the hypothesis that Pol IV and V genes resulted from a duplication event before the divergence of land plants. In addition, whereas *Arabidopsis* and other Brassicaceae species have one gene encoding the largest subunits of each of Pol II, IV and V, rice has two copies for each of the Pol II, IV and V largest subunits, with rice-specific *NRPB1* and *NRPD1* paralogs and *NRPE1* paralogs shared by grasses (Figs 4a, S10a), illustrating independent duplication, which also occurred in other angiosperm lineages (Fig. 4a).

The genes for the 2nd subunits of Pol II, IV and V form two clades (Figs 3, S1, S7): one for the eukaryotic Pol II 2nd subunits (*NRPB2*); the other with the *Arabidopsis NRPD2* gene encoding the shared Pol IV/V 2nd subunit and homologs from angiosperms, *Selaginella* and *Physcomitrella*. The lack of a *NRPD2* homolog in green algae and nonplant eukaryotes suggests that the *NRPD2* gene arose in early land plants and was maintained as a single copy for much of land plant history, but an earlier origin is possible. However, recent independent duplications have occurred in Brassicaceae and Poaceae: *Arabidopsis thaliana* has a second gene without a known function, as a result of a duplication shared by *Arabidopsis lyrata* but not other Brassicaceae species (Fig. S10b). Interestingly, *Zea mays* has five genes for the 2nd subunit: two (*NRPB2a*, *NRPB2b*) came from maize-specific duplication, whereas three *NRPD2* paralogs resulted from duplications in the common ancestor of grasses and in the common ancestor of maize and sorghum, respectively (Fig. S10b).

The 4th subunits of Pol I, II and III are encoded by three genes in yeast (Cramer *et al.*, 2008), but the plant homologs of the genes for Pol I and III have not been identified. Thus, only those related to the Pol II genes are analyzed here (Figs 3, S2). Two

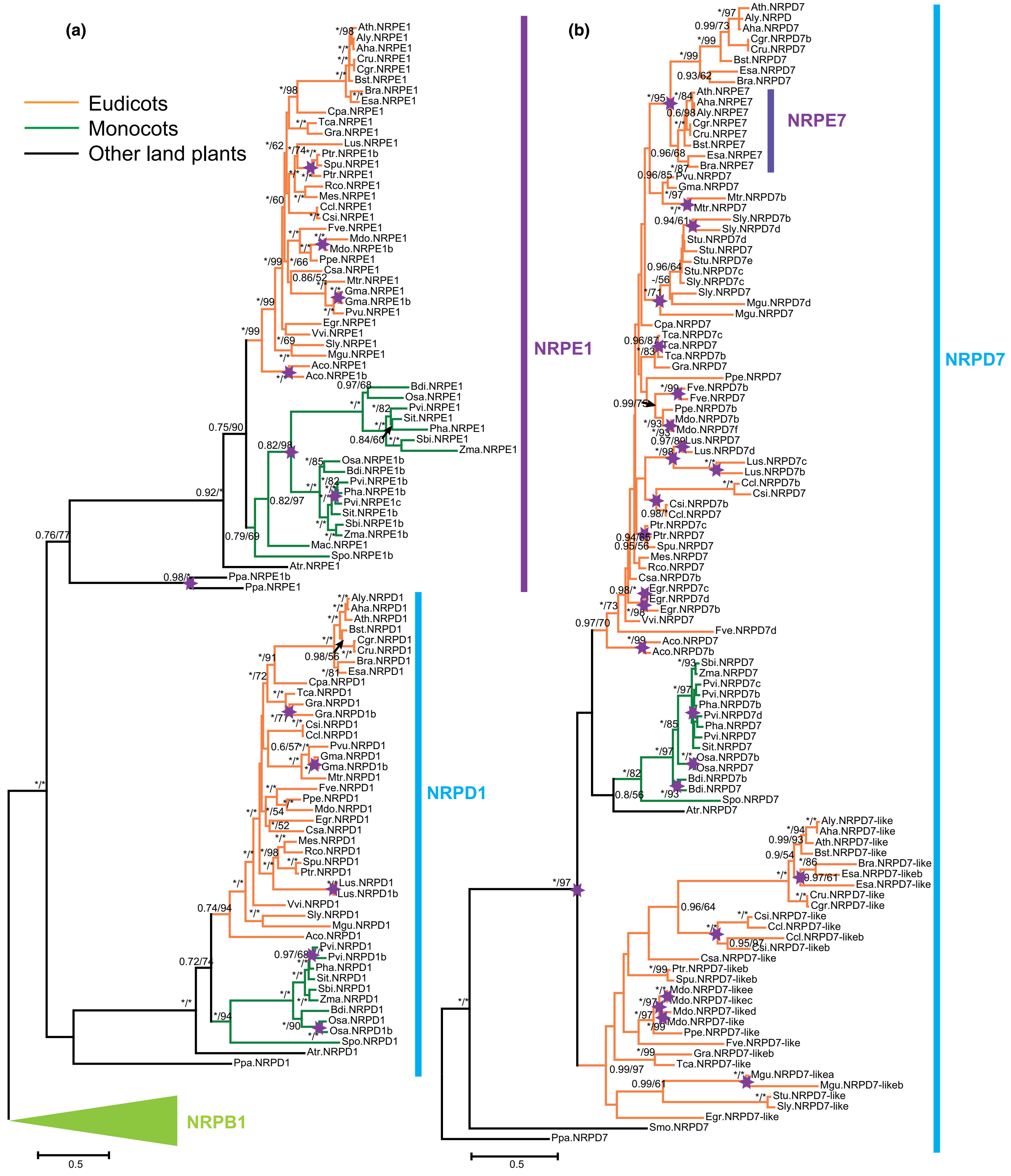


Fig. 4 Phylogenies of genes for representative Pol IV and Pol V subunits in plants: (a) 1st subunits; (b) 7th subunits. Tree topologies generated from RAxML are shown here. Bayesian posterior probability values (> 0.5) from MrBayes and bootstrap values (> 50) from RAxML are labeled on internal nodes. Asterisks (*) indicate Bayesian posterior probability values of 1 or bootstrap values of 100. Gene duplication events are marked with purple asterisks. Aco, *Aquilegia coerulea*; Aha, *Arabidopsis halleri*; Aly, *Arabidopsis lyrata*; Ath, *Arabidopsis thaliana*; Atr, *Amborella trichopoda*; Bdi, *Brachypodium distachyon*; Bra, *Brassica rapa*; Bst, *Boechera stricta*; Ccl, *Citrus clementina*; Cgr, *Capsella grandiflora*; Cpa, *Carica papaya*; Cru, *Capsella rubella*; Csa, *Cucumis sativus*; Csi, *Citrus sinensis*; Egr, *Eucalyptus grandis*; Esa, *Eutrema salsugineum*; Fve, *Fragaria vesca*; Gma, *Glycine max*; Gra, *Gossypium raimondii*; Lus, *Linum usitatissimum*; Mac, *Musa acuminata*; Mdo, *Malus domestica*; Mes, *Manihot esculenta*; Mgu, *Mimulus guttatus*; Mtr, *Medicago truncatula*; Osa, *Oryza sativa*; Pha, *Panicum hallii*; Ppa, *Physcomitrella patens*; Ppe, *Prunus persica*; Ptr, *Populus trichocarpa*; Pvi, *Panicum virgatum*; Pvu, *Phaseolus vulgaris*; Rco, *Ricinus communis*; Sbi, *Sorghum bicolor*; Sit, *Setaria italica*; Sly, *Solanum lycopersicum*; Smo, *Selaginella moellendorffii*; Spo, *Spirodela polyrhiza*; Spu, *Salix purpurea*; Stu, *Solanum tuberosum*; Tca, *Theobroma cacao*; Vca, *Volvox carteri*; Vvi, *Vitis vinifera*; Zma, *Zea mays*.

types of plant gene were identified: one for Pol II (*NRPB4*) and a second similar to the *Arabidopsis NRPD4* gene for Pol IV and V. The angiosperm *NRPB4* and *NRPD4* homologs form two sister clades, each containing sequences from eudicots and monocots, consistent with Tucker *et al.* (2011) using *Arabidopsis*, rice and maize genes. Further, we found that *Amborella* (a basal angiosperm) also has one gene in each clade, indicating that *NRPD4* genes probably resulted from duplication predating the diversification of the extant angiosperms. Analysis with the only detected *NRPB4* homolog from the gymnosperm Norway spruce suggests that this duplication occurred after angiosperms separated from gymnosperms.

In addition to the clades for Pol I and III, the gene family for the 7th subunits has one well-supported clade (Figs 3, S3) containing eukaryotic Pol II genes and land plant homologs of the *Arabidopsis NRPD7* and *NRPE7* genes. The latter clade contains two groups, one with the Pol II genes, and the other including the land plant Pol IV/V related genes, suggesting that the *NRPD7*/*NRPE7*-like genes originated from a duplicate copy of an ancestral *NRPB7* (Pol II) gene in early land plants, consistent with previous results (Tucker *et al.*, 2011). Further analysis with *NRPD7*/*NRPE7* homologs indicates that *Arabidopsis NRPD7* and *NRPE7* and their respective Brassicaceae orthologs resulted from duplication in the Brassicaceae common ancestor, after divergence from other eudicots (Figs 4b, 5a). Homologs of the ancestral *NRPD7*/*NRPE7* gene are found in other eudicots and monocots, as well as in nonflowering plants (Fig. 4b). Biochemical studies in maize indicated that Pol IV and Pol V shared the 7th subunit encoded by the only *NRPD7* gene in maize (Haag *et al.*, 2014) (Fig. 5a). Therefore, the 7th subunit has diverged between Pol IV and V in Brassicaceae, but not in maize. Many independent duplications of the gene for the Pol IV/V 7th subunit were detected here (Fig. 4b), potentially allowing functional divergence between Pol IV and V.

The 9th subunit genes form three clades for eukaryotic Pol I, II and III, respectively (Figs 3, S4). In each clade, *Arabidopsis* has two copies, probably resulting from duplication after the divergence of Brassicaceae and other eudicots (Fig. S8). Despite the highly similar sequences and presence in both Pol IV and V (Ream *et al.*, 2009), genetic analysis in *Arabidopsis* revealed that *NRPB9a* and *NRPB9b* are functionally different, with only the *rpb9b* mutant showing defects in RNA-directed DNA methylation (Tan *et al.*, 2012). Grasses have three groups of *NRPB9* genes, implying at least two rounds of duplication in their ancestor (Fig. S10e).

Type β: genes for the 3rd and 11th subunits The families for both the 3rd and 11th subunits each contain two ancient groups (Figs 2b, 3, S5). The first group (*NRPA3* for the 3rd subunit and *NRPA11* for the 11th subunit) is for both eukaryotic Pol I and III, whereas the second group (*NRPB3* and *NRPB11*) contains genes for Pol II subunits. Like animals and fungi, most plant species have one copy in each group, but *Arabidopsis* has two copies for each of *NRPA3* and *NRPB3*, and rice has two copies of *NRPA11* and *NRPA3* genes. Phylogenetic analysis of Brassicaceae genes revealed that *NRPA3* genes underwent at least one duplication in the Brassicaceae common ancestor (Fig. S11a). Grass gene trees showed a rice-specific duplication for two *NRPA3* genes, but an ancestral grass duplication for the *NRPA11* genes (Fig. S11a,b).

Type γ: genes for the 5th, 6th, 8th, 10th and 12th subunits The genes encoding the 5th subunits of all RNAPs are single copy in most animals, fungi, green algae and *Selaginella*, but have multiple copies in angiosperms (Figs 1, S9). The angiosperm genes form two clades (Figs 2c, S9): one includes genes encoding for the 5th subunit shared by Pol I, II, III and IV in *Arabidopsis*, and the other, with longer branch lengths, contains genes for the Pol V 5th subunit, suggesting that a gene duplication event occurred before the diversification of angiosperms. Genes in both angiosperm clades underwent more independent duplications after eudicots and monocots diverged (Figs S9, 11c). Within the clade of genes for *Arabidopsis* Pol I–IV, Brassicaceae genes formed two monophyletic groups, implying gene duplication in their ancestor (Fig. S11c). These results are in agreement with the previous analysis of *Arabidopsis*, rice and maize genes (Tucker *et al.*, 2011). Our analysis of the monocot genes included those of banana and six grasses, and the results indicate that the duplication occurred before Poales split from Zingiberales (Fig. S11c). In maize, biochemical analyses showed that both *NRPB5* genes encode the 5th subunit of Pol II (Haag *et al.*, 2014). In the second clade, eudicot genes underwent gene duplication events, resulting in two sister groups; one of these duplicates experienced a subsequent duplication event after the divergence of Brassicaceae and other eudicots, consistent with a whole-genome duplication (WGD) event in the common ancestor of Brassicaceae. (Figs S9, S11c).

In addition, phylogenetic trees revealed that the genes for the 6th, 8th and 12th subunits underwent duplications in the common ancestor of grasses, and the genes for the 6th, 8th, 10th and 12th subunits underwent duplications in the Brassicaceae common ancestor (Figs S6, S11d–g), revealing more lineage-specific gene duplications.

Expression and duplication patterns of RNAP subunits

On the basis of earlier described phylogenetic analyses, we reconstructed the evolutionary history of all the genes related to RNAP subunits in *Arabidopsis* (Figs 5, S12a) and maize (Figs 5, S13a). Although having experienced relatively short histories, pairs of recent duplicates in these two species, including *Arabidopsis NRPB3a*/*NRPB3b*, *NRPE5*/*NRPD5b*, *NRPD7*/*NRPE7*, *NRPB8a*/*NRPB8b*, *NRPB9a*/*NRPB9b*, *NRPB10*/*NRPB10b* and *NRPB12*/*NRPB12b*, and maize *NRPE1a*/*NRPE1b*, *NRPD2a*/*NRPD2b*/*NRPE2c*, *NRPB5a*/*NRPB5b*, *NRPB8*/*NRPB8b*, *NRPB9a*/*NRPB9b* and *NRPB10*/*NRPB10b*, exhibit differential gene expression patterns (Figs S12b, S13b) and distinct properties in terms of participation or not in relevant protein complexes or the extent of such participation (Fig. S12c). For example, one type of 12th subunit (NRPB12a) was detected by mass spectrometry using purified polymerases from *Arabidopsis* (Ream *et al.*, 2009); nevertheless, the second copy (*NRPB12b*) was expressed

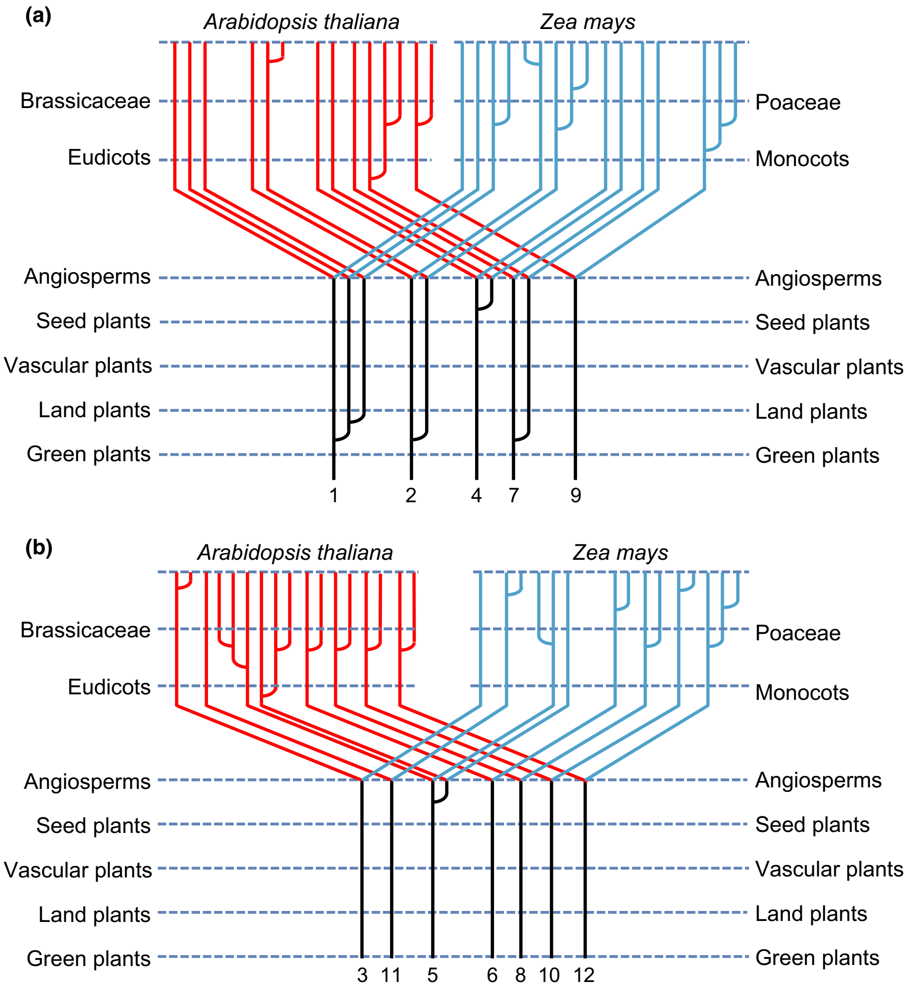


Fig. 5 Evolutionary histories of Pol II, Pol IV and Pol V subunits in *Arabidopsis* and maize (*Zea mays*). (a) Type α subunits. (b) Type β and γ subunits. Evolutionary histories in eudicots and monocots are colored in red and blue, respectively.

at higher levels in seedlings and leaves than in most tissues, suggesting that NRPB12b might function in specific tissues. These findings suggest that these subunits have undergone functional diversification.

To further examine parallel duplications of RNAPs, focusing on the subunits that have experienced diversification during plant evolution, we analyzed the gene duplication patterns in Brassicaceae and Poaceae by reconciling gene trees with species trees. For 28 genes encoding such subunits, we identified multiple independent duplications in Fig. 6. The duplication and loss patterns are different between Brassicaceae and Poaceae, as well as among different subunits, suggesting that RNAPs might have evolved different functions in separate plant lineages. Brassicaceae underwent two rounds of WGDs after separating from other eudicots. Poaceae also underwent two WGDs after divergence from other monocots. However, for 28 genes present in the shared ancestor of both families, only nine duplications were detected in the separate ancestors of each family, respectively, suggesting that most RNAP genes were lost after WGD, but before the divergence of the species examined here. In addition, many duplications were species specific, especially in *Brassica rapa* and *Panicum virgatum*, probably because each of these plants underwent additional independent WGD(s).

Sequence analysis suggests subunit coevolution within an RNAP complex and divergence between different RNAPs

To compare selection pressures on different subunits of different RNAPs at the same timescale, we reconstructed ancestral sequences of each gene in the common ancestor of Brassicaceae and calculated the ratio of the rates of nonsynonymous to synonymous substitutions ($\omega = d_N/d_S$) between each Brassicaceae gene and the ancestral gene. Similar analyses were performed for Poaceae genes. Interestingly, genes encoding different subunits of the same RNAP have similar d_N/d_S ratios; by contrast, homologs encoding the same subunit for different RNAPs exhibit different ω values (Fig. 7), with specific values presented for each species (Figs 7, S14a,b). The results strongly support the hypothesis that a protein complex can be considered as an evolutionary unit whose components coevolve under similar selection pressure. On comparison with genes for Pol I- and Pol III-specific subunits, genes for Pol II subunits have the lowest d_N/d_S, ratios, in plants (Fig. 7) as well as in animals and fungi (Fig. S14c,d), indicating that genes for Pol II subunits are under strong purifying selection. Genes for Pol IV and Pol V subunits have much higher d_N/d_S ratios than their Pol II homologs (Fig. 7). Our results are consistent with and extend beyond the previous analysis using the 1st

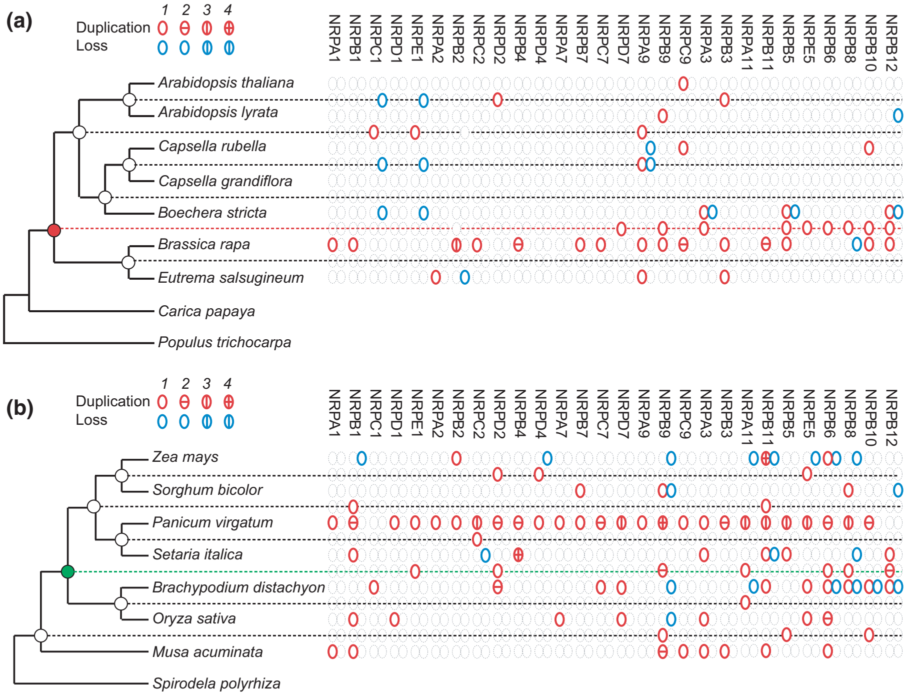


Fig. 6 Gene duplication and loss patterns of RNA polymerase (RNAP) subunit genes with variable copy numbers. Duplication and loss events are indicated by red and blue ovals, respectively, for each species and ancestral node, in comparison with the previous node, on the phylogenies of (a) Brassicaceae and (b) Poaceae.

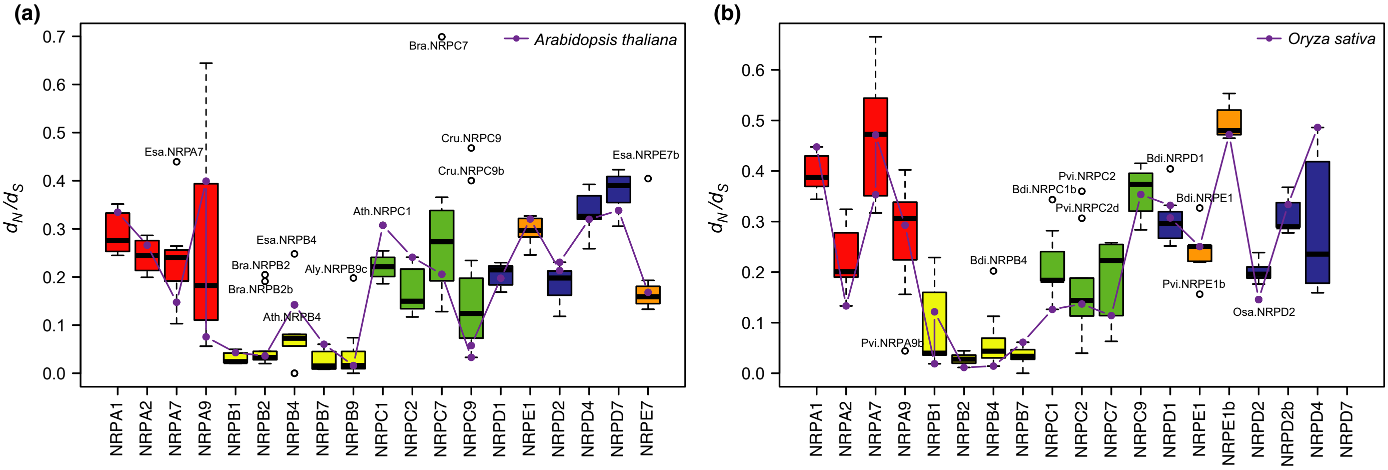


Fig. 7 Selection pressure on genes for RNA polymerase (RNAP) subunits. The distributions of nonsynonymous to synonymous rate (d_N/d_S) ratios (ω) of type α subunits in (a) Brassicaceae and (b) Poaceae are shown in the boxplots. Genes with extreme values (outside of the 1.5 interquartile range) are indicated by circles. Pol I-, Pol III- and Pol V-specific subunits are colored in red, green and orange, respectively. Pol II-specific and Pol II, Pol IV and Pol V shared subunits are colored in yellow. Pol IV-specific and Pol IV and Pol V shared subunits are colored in blue. The d_N/d_S ratios of genes from (a) *Arabidopsis thaliana* and (b) *Oryza sativa* are indicated with purple circles and lines.

and 2nd subunits from *Arabidopsis* and rice (Luo & Hall, 2007). Brassicaceae *NRPD7* and *NRPE7* originated from duplication in the common ancestor of Brassicaceae and diverged under different selection pressures (Fig. 7a). Two copies of *NRPE1* and *NRPD2* in Poaceae also have distinct d_N/d_S ratios (Fig. 7b), suggesting different evolutionary constraints after duplication. In addition, some recent duplicates in both families, such as Bra.NRPB2 and Bra.NRPB2b and Pvi.NRPC2 and Pvi.NRPC2d, showed much greater ω values than the average of the same subunit in the same RNAP, suggesting that these genes are less constrained and possible candidates for new functions.

Origin of the largest subunit of plant-specific RNAPs

The Pol IV/V largest subunits contain C-terminal regions that are different from those of Pol I–III. Sequence analysis showed that the Pol IV/V C-terminal region is conserved among plants and contains Domain of Unknown Function 3223 (DUF3223) as defined in the Pfam database. DUF3223 is detected in plants, and nonpolymerase proteins in bacteria and protists, but not in animals and fungi. To investigate further the distribution of DUF3223, we searched plant genomes for genes encoding proteins with DUF3223, and found five in each of *Arabidopsis*, rice and *Amborella*, seven in *Physcomitralla* and one in green algae (Fig. 8). The *Arabidopsis* genes are *NRPD1*, *NRPE1*, *DeCL* (*DEFECTIVE CHLOROPLASTS AND LEAVES*), *DeCL-like* and *Domino1*.

Phylogenetic analysis of plant genes encoding proteins with DUF3223 showed that *NRPD1* and *NRPE1* genes are clustered with *DeCL* and *DeCL-like* genes with high support values, and contain sequences from the land plant species examined here (Fig. 8). Gene structure analysis showed that the DUF3223 domains of *NRPD1* and *NRPE1* are encoded by three exons, separated by 'phase 2' and 'phase 0' introns. Strikingly, *DeCL* and *DeCL-like* genes also have three exons with the 'phase 2' first intron and the 'phase 0' second intron, congruent with introns within the *NRPD1/NRPE1* DUF3223 domains (Figs 8, S15a). Both phylogenetic results and gene structure conservation support the idea that the DUF3223 domains of *NRPD1*/*NPRE1* and *DeCL*/*DeCL-like* genes have a common ancestor in land plants. As *NRPD1*/*NRPE1* have N-terminal sequences homologous to *NRPB1* and C-terminal sequences homologous to *DeCL*/*DeCL-like* genes, *NRPD1* and *NRPE1* could have resulted from a gene fusion event of an *NRPB1-like* gene and a *DeCL-like* gene in early land plants (Fig. S15b).

Discussion

Duplications of RNAP subunits at different times in land plant history and independently in angiosperm groups

We have performed a comprehensive evolutionary analysis of RNAP gene families. Our results and previous findings of Pol IV/V subunit compositions in *Arabidopsis* (Huang *et al.*, 2009; Ream *et al.*, 2009; Law *et al.*, 2011) (Fig. S12c) and maize (Haag *et al.*, 2014) (Fig. S13c) support a multistep model for the evolution of plant RNAPs (Fig. 9b). In this model, plant-specific subunits were derived from eukaryotic homologs at different times, resulting in increased RNAP diversity. As first proposed by Luo & Hall (2007), an early event was the duplication of *NRPB1* in the ancestor of land plants and the Charophyta, although an even earlier origin is possible. Subsequently, a fusion of one duplicate with a *DeCL-like* gene produced the common ancestral gene for *NRPD1* and *NRPE1*. This ancestral gene then duplicated to generate *NRPD1* and *NRPE1* in the common land plant ancestor. The genes for the Pol II 2nd and 7th subunits (*NRPB2* and *NRPB7*) also duplicated early in land plants, resulting in one new copy for each gene, encoding the Pol IV/V 2nd and 7th subunits, respectively. In short, the land plant ancestor should have had the NRPD1–NRPD2–NRPD7 and NRPE1–NRPD2–NRPD7 combinations, whereas the other subunits were probably shared by Pol II, IV and V.

During angiosperm evolution, further independent duplications and differentiations resulted in greater divergence of subunit composition between Pol II and IV/V and among different angiosperm groups. In the angiosperm ancestor, additional duplication yielded new genes for the Pol IV/V 4th subunits (*NRPD4*) and the Pol V 5th subunit (*NRPE5*). After the divergence of monocots and eudicots, RNAP genes underwent separate duplications, resulting in different evolutionary patterns in these two major angiosperm groups. Within each group, genes for several subunits further duplicated independently in the ancestors of Brassicaceae and Poaceae, increasing the complexity for subunit compositions. Some duplicates have not yet differentiated functionally and are shared between Pol II, IV and V (e.g. multiple

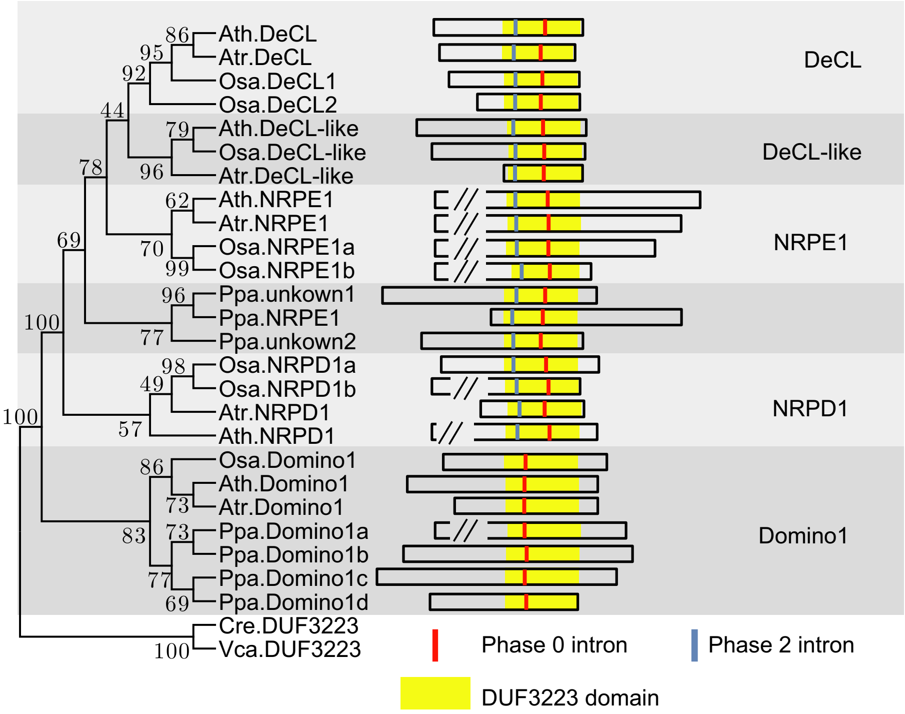


Fig. 8 A proposed origin of the Domain of Unknown Function 3223 (DUF3223) in the 1st subunits of Pol IV and Pol V. A maximum likelihood (ML) tree of plant genes encoding proteins with a DUF3223 domain. Structures of DUF3223 domain genes are shown following each gene name. The DUF3223 domain is colored in yellow. Introns located in the region of the DUF3223 domain are marked with red (phase 0 introns) or blue (phase 1 introns) vertical lines. Ath, *Arabidopsis thaliana*; Atr, *Amborella trichopoda*; Cre, *Chlamydomonas reinhardtii*; Osa, *Oryza sativa*; Ppa, *Physcomitrella patens*; Vca, *Volvox carteri*.

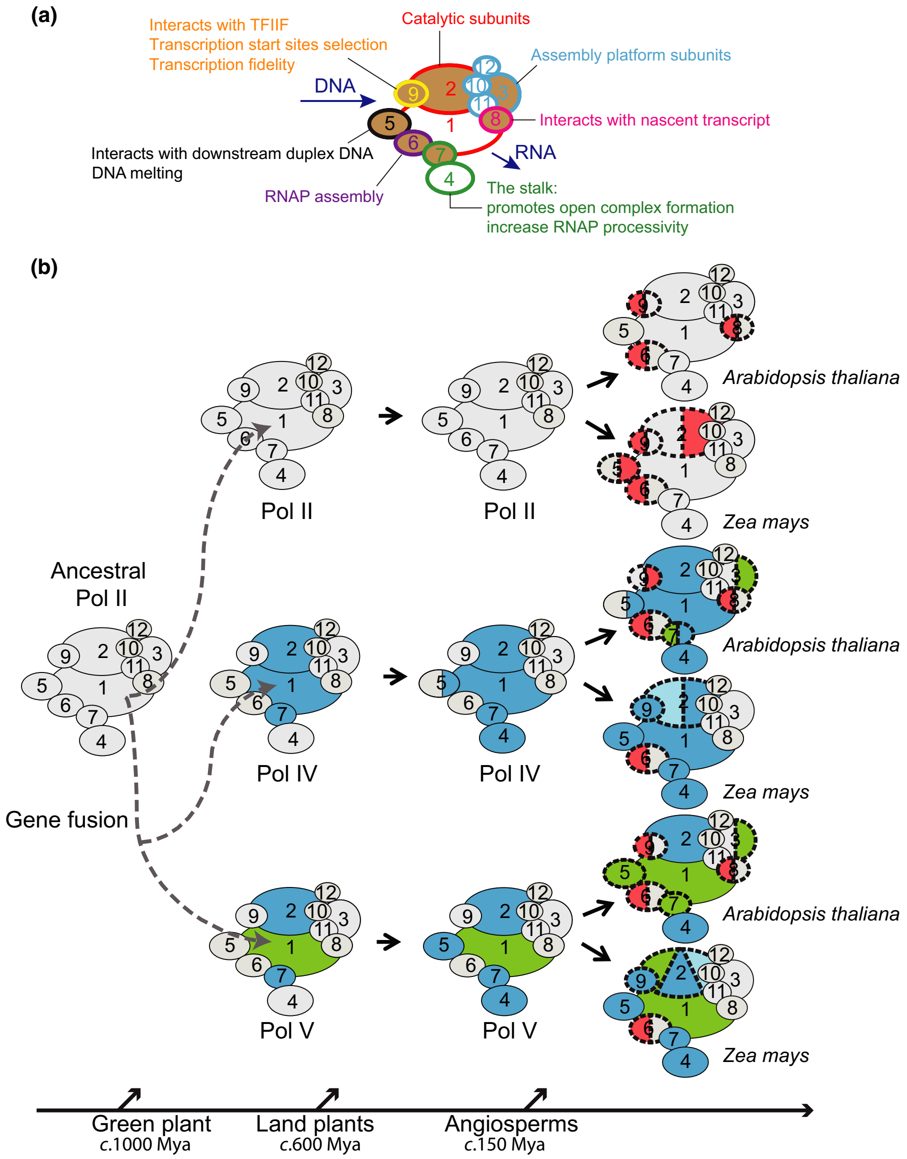


Fig. 9 Pol II, IV and V subunit compositions during plant evolution. (a) Functions of Pol II subunits in yeast summarized from previous studies (Minakhin *et al.*, 2001; Werner & Grohmann, 2011). Subunits with different compositions in Pol II, IV and V, between *Arabidopsis* and maize (*Zea mays*), are shown in brown. (b) Pol II, IV and V subunit compositions during plant evolution. Subunits encoded by the same gene for Pol II, IV and V are indicated in gray. Subunits encoded by duplicate genes are shown in red, blue and green, respectively. Subunits with alternative copies are shown in two or three colors. The duplication and gene fusion events of the largest subunits are indicated by dashed lines on the left. The independently duplicated subunits in *Arabidopsis* and maize after the split of eudicots and monocots are indicated by dashed circles on the right. The subunit compositions of Pol II, IV and V in *Arabidopsis* and maize were from previous biochemical studies (Huang *et al.*, 2009; Ream *et al.*, 2009; Law *et al.*, 2011; Haag *et al.*, 2014).

NRPB9 in *Arabidopsis* and multiple NRPB6 in both *Arabidopsis* and maize). By contrast, the genes for the 7th subunit, duplicated in the ancestor of Brassicaceae, have differentiated for Pol IV and V, respectively (Ream *et al.*, 2009). *NRPD2*/*NRPE2* duplicated in the common ancestor of Poaceae, and then duplicated further in *Z. mays*, resulting in two copies of the 2nd subunit for Pol IV and three isoforms of the 2nd subunit for Pol V (Haag *et al.*, 2014). Other independent duplications have occurred in different angiosperm groups (Figs 4, S7–S9); further biochemical studies in different plants might reveal even greater functional diversity among RNAPs.

A crucial role of WGD in RNAP diversification

RNAP evolution is possibly facilitated by numerous genome duplications. WGDs occur widely in animals, fungi and plants (Kellis *et al.*, 2004; Kasahara, 2007; Jiao *et al.*, 2011; Lee *et al.*, 2013) and have been implicated in the evolution of novel complexes and pathways because of the simultaneous duplication of all components (Jaillon *et al.*, 2009). However, the functional differentiation of components of a complex after WGD is much less well understood. The evolution of beneficial novel function in one subunit might be constrained by necessary interactions with other subunits. In plants, WGDs are particularly prevalent (Jiao *et al.*, 2011; Lee *et al.*, 2013) and might have played an important role in RNAP evolution. Indeed, phylogenetic analyses of RNAP subunit genes placed duplication events at the same node as known plant WGDs (Jiao *et al.*, 2011; Lee *et al.*, 2013), including those in the angiosperm ancestor (4th and 5th subunits; Figs S2, S9), eudicot ancestor (5th and 7th subunits; Figs 4b, S9), Brassicaceae ancestor (3rd, 5th, 6th, 7th, 8th, 9th, 10th and 12th subunits; Fig. 6a) and grass ancestor (1st, 2nd, 6th, 8th, 9th, 11th and 12th subunits; Fig. 6b). For many recent duplicate gene pairs, strong evidence for duplication as a result of WGD was detected for the RNAP genes in syntenic regions from WGDs (Figs S10, S11).

Although each WGD duplicates the entire genome, we found that only a small fraction of genes is retained subsequently and gene retention rates vary among different subunits (Fig. 6), possibly because of different functional or structural constraints on different subunits. It is possible that RNAPs might have functionally diverse forms among different plants. For example, only one *NRPE1* gene was detected in the *Arabidopsis* genome, but two sets of *NRPE1* genes were found in grass genomes (*NRPE1* and *NRPE1b*). *NRPE1b* evolved rapidly with relatively higher d_N/d_S values than any other RNAP gene, and the maize *NRPE1b* is specifically expressed in immature tassels, suggesting that *NRPE1b* might have a novel function associated with reproductive development in grasses. Both *Arabidopsis* and maize have multiple copies of *NRPD2* genes. Expression and genetic analysis showed that only one of the two *Arabidopsis NRPD2* genes has detectable function (Onodera *et al.*, 2005), but all three maize *NRPD2* paralogs are expressed and have distinct functions (Sidorenko *et al.*, 2009; Stonaker *et al.*, 2009). Interestingly, maize *NRPE2c* is preferentially expressed in the tassel and pollen. These findings suggest that grasses might possess distinct types of RNAPs compared with those of *Arabidopsis*.

Distinct evolutionary patterns are associated with diverse subunit functions

The heteromeric RNAPs have at least 12 subunits with different functional and structural characteristics (Fig. 9a), which could explain the various evolutionary patterns for different subunits presented here. This is consistent with previous studies showing that different subunits from the same complex have discordant evolutionary patterns (Matalon *et al.*, 2014). Plants and other eukaryotes have distinct 1st, 2nd, 4th, 7th and 9th subunits for Pol I, II and III. During plant evolutionary history, genes for the 1st, 2nd, 4th and 7th subunits evolved novel copies for plant-specific subunits of Pol IV and V. The 1st and 2nd subunits are catalytic subunits, responsible for DNA binding and RNA synthesis (Cramer *et al.*, 2001, 2008). In Pol II, the 4th and 7th subunits form a subcomplex that interacts with the two largest subunits and is involved in recruitment of 3′-processing factors and mRNA export (Runner *et al.*, 2008; Harel-Sharvit *et al.*, 2010). The duplicate copies of these subunits could have contributed to RNAP functional diversity by allowing the optimization of interactions for different DNA templates and RNA products. Although the gene tree for the 9th subunit has no plant-specific clade, *NRPB9* underwent duplication in the Brassicaceae ancestor and two rounds of duplication in the grass ancestor, with the *Arabidopsis* paralogs having partially redundant functions in Pol IV/Pol V pathways (Tan *et al.*, 2012) and maize having two and one paralogs specific for Pol II and Pol IV/Pol V, respectively. Therefore, compared with the other subunits distinct for different types of RNAPs, the Pol IV/V 9th subunit diverged from Pol II most recently.

Proteins with multiple interactive partners via the same interactive site are highly conserved because sequence or copy number changes can affect many interactions, whereas proteins interacting with different partners using different interactive sites can coevolve with the site-specific partner. The evolutionary patterns of eukaryotic RNAP common subunits (6th, 8th, 10th and 12th) partly support this hypothesis: common subunits have more conserved sequences and generally fewer copies; only very recent duplicate copies are retained; the duplicates might be in the process of functional divergence. However, the gene family for the 5th subunit is an exception. Only one 5th subunit gene is found in animal or fungal genomes, but it has duplicated multiple times in plant history. In *Arabidopsis*, NRPB5 is shared by Pol I, II, III and IV and NRPE5 is specific for Pol V (Ream *et al.*, 2009). Structural studies of the yeast RNAPs showed that RPB5 contacts DNA ahead of a transcriptional fork with its N-terminal region binding to RPB1 and C-terminal region interacting with transcription factors (Cramer *et al.*, 2001, 2008). The novel 5th subunits in plants might be responsible for the recognition of different DNA templates and binding to transcriptional factors for Pol V-specific transcription. The new 5th subunit gene (*NRPE5*) arose in angiosperms, suggesting that it is important for an epigenetic mechanism specific to flowering plants.

Our analysis further suggests that recently duplicated RNAP genes tend to duplicate further, such as *NRPE1*, *NRPD2*, *NRPE5* and *NRPD7*. One plausible explanation is that these genes are under reduced selection pressure. These genes have higher d_N/d_S ratios, indicating that they are under weaker purifying selection (Fig. 7). In addition, mutants of *NRPE1* and *NRPE5* have less defective phenotypes than mutants of their counterparts in Pol I, II and III (Onodera *et al.*, 2005, 2008; Lahmy *et al.*, 2009; Ream *et al.*, 2009), consistent with the idea that newly arising subunits are under reduced evolutionary constraints. Our results support a step-wise model for the evolution of multisubunit RNAPs in plants, with independent duplication and diversification in different plant groups, providing insights into the evolution of multisubunit protein complexes in general.

Acknowledgements

We would like to thank Yazhou Sun for providing Javascript for the prediction of homologs from genome sequences. We are grateful to Ji Qi, Hongzhi Kong, Ji Yang, Ning Zhang, Liping Zeng, Haifeng Wang, Qi Li, Ren, Guojun Zhou, Libing Shen and Jiemeng Liu for helpful discussions and comments on the manuscript and Lei Wang for providing the computer server. This work was supported by grants from the National Natural Science Foundation of China (91131007) and the Ministry of Science and Technology of China (2011CB944600) and by funds from the State Key Laboratory of Genetic Engineering.

References

Archambault J, Friesen JD. 1993. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiological Reviews* 57: 703–724.

Baker CR, Hanson-Smith V, Johnson AD. 2013. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* 342: 104–108.

Beck F, Unverdorben P, Bohn S, Schweitzer A, Pfeifer G, Sakata E, Nickell S, Plitzko JM, Villa E, Baumeister W *et al.* 2012. Near-atomic resolution structural model of the yeast 26S proteasome. *Proceedings of the National Academy of Sciences, USA* **109**: 14870–14875.

Ben-Shem A, Garreau de Loubresse N, Melnikov S, Jenner L, Yusupova G, Yusupov M. 2011. The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**: 1524–1529.

Boyer PD. 1997. The ATP synthase – a splendid molecular machine. *Annual Review of Biochemistry* **66**: 717–749.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology* **7**: 429–447.

Cramer P, Armache KJ, Baumli S, Benkert S, Brueckner F, Buchen C, Damsma GE, Dengl S, Geiger SR, Jasiak AJ *et al.* 2008. Structure of eukaryotic RNA polymerases. *Annual Review of Biophysics* **37**: 337–352.

Cramer P, Bushnell DA, Kornberg RD. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**: 1863–1876.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**: 1164–1165.

Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA. 2012. PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Research* **40**: D1194–D1201.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Erhard KF Jr, Parkinson SE, Gross SM, Barbour JE, Lim JP, Hollick JB. 2013. Maize RNA polymerase IV defines trans-generational epigenetic variation. *Plant Cell* **25**: 808–819.

Erhard KF Jr, Stonaker JL, Parkinson SE, Lim JP, Hale CJ, Hollick JB. 2009. RNA polymerase IV functions in paramutation in *Zea mays*. *Science* **323**: 1201–1205.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

Haag JR, Brower-Toland B, Krieger EK, Sidorenko L, Nicora CD, Norbeck AD, Irsigler A, LaRue H, Brzeski J, McGinnis K *et al.* 2014. Functional diversification of maize RNA polymerase IV and V subtypes via alternative catalytic subunits. *Cell Report* **9**: 378–390.

Haag JR, Pikaard CS. 2011. Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature Reviews. Molecular Cell Biology* **12**: 483–492.

Harel-Sharvit L, Eldad N, Haimovich G, Barkai O, Duek L, Choder M. 2010. RNA polymerase II subunits link transcription and mRNA decay to translation. *Cell* **143**: 552–563.

Huang L, Jones AM, Searle I, Patel K, Vogler H, Hubner NC, Baulcombe DC. 2009. An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nature Structural & Molecular Biology* **16**: 91–93.

Hull MW, Mckune K, Woychik NA. 1995. RNA polymerase II subunit Rpb9 is required for accurate start site selection. *Genes & Development* **9**: 481–490.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews. Genetics* **11**: 97–108.

Jaillon O, Aury JM, Wincker P. 2009. "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. *Comptes Rendus Biologies* **332**: 241–253.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kasahara M. 2007. The 2R hypothesis: an update. *Current Opinion in Immunology* **19**: 547–552.

Kelleher RJ 3rd, Flanagan PM, Kornberg RD. 1990. A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell* **61**: 1209–1215.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.

Kelman Z, O'Donnell M. 1995. DNA polymerase III holoenzyme: structure and function of a chromosomal replicating machine. *Annual Review of Biochemistry* **64**: 171–200.

Kolde R. 2013. *pheatmap: pretty heatmaps*. R package version 0.7.7.

Lahmy S, Pontier D, Cavel E, Vega D, El-Shami M, Kanno T, Lagrange T. 2009. PolV(PolIVb) function in RNA-directed DNA methylation requires the conserved active site and an additional plant-specific subunit. *Proceedings of the National Academy of Sciences, USA* **106**: 941–946.

Laubinger S, Zeller G, Henz SR, Sachsenberg T, Widmer CK, Naouar N, Vuylsteke M, Scholkopf B, Ratsch G, Weigel D. 2008. At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in *Arabidopsis thaliana*. *Genome Biology* **9**: R112.

Law JA, Vashisht AA, Wohlschlegel JA, Jacobsen SE. 2011. SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS Genetics* **7**: e1002195.

Lee TH, Tang H, Wang X, Paterson AH. 2013. PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research* **41**: D1152–D1158.

Lin Z, Kong H, Nei M, Ma H. 2006. Origins and evolution of the *recA*/*RAD51* gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proceedings of the National Academy of Sciences, USA* **103**: 10328–10333.

Lin Z, Nei M, Ma H. 2007. The origins and early evolution of DNA mismatch repair genes – multiple horizontal gene transfers and co-evolution. *Nucleic Acids Research* **35**: 7591–7603.

Luo J, Hall BD. 2007. A multistep process gave rise to RNA polymerase IV of land plants. *Journal of Molecular Evolution* **64**: 101–112.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Matalon O, Horovitz A, Levy ED. 2014. Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. *Current Opinion in Structural Biology* **26**: 113–120.

Minakhin L, Bhagat S, Brunning A, Campbell EA, Darst SA, Ebright RH, Severinov K. 2001. Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proceedings of the National Academy of Sciences, USA* **98**: 892–897.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology* **8**: 122–128.

Neer EJ. 1995. Heterotrimeric G proteins: organizers of transmembrane signals. *Cell* **80**: 249–257.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nature Reviews. Genetics* **9**: 951–963.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**: 121–152.

Onodera Y, Haag JR, Ream T, Costa Nunes P, Pontes O, Pikaard CS. 2005. Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**: 613–622.

Onodera Y, Nakagawa K, Haag JR, Pikaard D, Mikami T, Ream T, Ito Y, Pikaard CS. 2008. Sex-biased lethality or transmission of defective transcription machinery in Arabidopsis. *Genetics* **180**: 207–218.

Pikaard CS, Haag JR, Pontes OM, Blevins T, Cocklin R. 2012. A transcription fork model for Pol IV and Pol V-dependent RNA-directed DNA methylation. *Cold Spring Harbor Symposia on Quantitative Biology* **77**: 205–212.

Pikaard CS, Tucker S. 2009. RNA-silencing enzymes Pol IV and Pol V in maize: more than one flavor? *PLoS Genetics* **5**: e1000736.

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution* **17**: 890–896.

R Core Team. 2014. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [WWW.document] URL http://www.R-project.org/ [accessed 21 April 2015].

Ream TS, Haag JR, Wierzbicki AT, Nicora CD, Norbeck AD, Zhu JK, Hagen G, Guilfoyle TJ, Pasa-Tolic L, Pikaard CS. 2009. Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Molecular Cell* **33**: 192–203.

Roeder RG, Rutter WJ. 1969. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* 224: 234–237.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61: 539–542.

Runner VM, Podolny V, Buratowski S. 2008. The Rpb4 subunit of RNA polymerase II contributes to cotranscriptional recruitment of 3′ processing factors. *Molecular and Cellular Biology* 28: 1883–1891.

Sampath V, Sadhale P. 2005. Rpb4 and Rpb7: a sub-complex integral to multi-subunit RNA polymerases performs a multitude of functions. *IUBMB Life* 57: 93–102.

Sidorenko L, Dorweiler JE, Cigan AM, Arteaga-Vazquez M, Vyas M, Kermicle J, Jurcin D, Brzeski J, Cai Y, Chandler VL. 2009. A dominant mutation in mediator of paramutation2, one of three-second-largest subunits of a plant-specific RNA polymerase, disrupts multiple siRNA silencing processes. *PLoS Genetics* 5: e1000725.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stonaker JL, Lim JP, Erhard KF Jr, Hollick JB. 2009. Diversity of Pol IV function is defined by mutations at the maize *rmr7* locus. *PLoS Genetics* 5: e1000706.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* 34: W609–W612.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 30: 2725–2729.

Tan EH, Blevins T, Ream TS, Pikaard CS. 2012. Functional consequences of subunit diversity in RNA Polymerases II and V. *Cell Report* 1: 208–214.

Tucker SL, Reece J, Ream TS, Pikaard CS. 2011. Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. *Cold Spring Harbor Symposia on Quantitative Biology* 75: 285–297.

Vannini A, Cramer P. 2012. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular Cell* 45: 439–446.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H *et al.* 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 40: e49.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.

Werner F, Grohmann D. 2011. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews. Microbiology* 9: 85–98.

Xu G, Ma H, Nei M, Kong H. 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proceedings of the National Academy of Sciences, USA* 106: 835–840.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology Evolution* 17: 32–43.

Zhou X, Ma H. 2008. Evolutionary history of histone demethylase families: distinct evolutionary patterns suggest functional divergence. *BMC Evolutionary Biology* 8: 294.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 The phylogeny of genes for the 2nd subunits in plants, animals and fungi.

Fig. S2 The phylogeny of genes for the 4th subunits in plants, animals and fungi.

Fig. S3 The phylogeny of genes for the 7th subunits in plants, animals and fungi.

Fig. S4 The phylogeny of genes for the 9th subunits in plants, animals and fungi.

Fig. S5 The phylogeny of genes for the 11th subunits in plants, animals and fungi.

Fig. S6 The phylogenies of genes for type γ RNA polymerase (RNAP) subunits.

Fig. S7 The phylogeny of genes for the 2nd subunits of Pol II, Pol IV and Pol V.

Fig. S8 The phylogeny of genes for the 9th subunits of Pol II, Pol IV and Pol V.

Fig. S9 The phylogeny of genes for the 5th subunits of Pol II, Pol IV and Pol V.

Fig. S10 The phylogenies of genes for type α RNA polymerase (RNAP) subunits in Brassicaceae and Poaceae.

Fig. S11 The phylogenies of genes for type β and type γ RNA polymerase (RNAP) subunits in Brassicaceae and Poaceae.

Fig. S12 Evolutionary and functional diversification of RNA polymerase (RNAP) subunits in *Arabidopsis.*

Fig. S13 Evolutionary and functional diversification of RNA polymerase (RNAP) subunits in *Zea mays.*

Fig. S14 Selection pressure on type α subunits in plants, animals and fungi.

Fig. S15 Evolution of the 1st subunits of Pol IV and Pol V.

Table S1 Information on genes for RNA polymerase (RNAP) subunits in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*

Table S2 Information on the sequence databases used in this study

Table S4 Best-fit models used in the phylogenetic analysis

Table S3 List of genes for the RNA polymerase (RNAP) subunits included in this study

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.