The evolution and composition of RNA polymerase IV in plants

Jie Luo

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree: Department of Biology

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Jie Luo

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Benjamin D. Hall

Reading Committee:

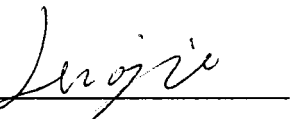
Benjamin D. Hall

Elizabeth van Volkenburgh

Willie J. Swanson

Date: 04/08/06

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform.

Signature 

Date 04/08/06

University of Washington

Abstract

The Evolution and Composition of RNA Polymerase IV in Plants
Jie Luo

Chair of the Supervisory Committee:
Professor Benjamin D. Hall
Department of Genome Sciences and Biology

This thesis explores two aspects of the evolution of RNA polymerase in plants. In both cases, components of the RNA polymerase II transcription system of the nucleus were found to have undergone duplicative proliferation.

In the derived angiosperm taxa *Ericales* and *Lamiidae*, two forms exist for the gene RPB2, encoding the second-largest subunit of RNA polymerase II. Other angiosperms possess only a single form. Chapter 2 of this thesis shows that the two paralogous forms of RPB2 evolved as separate lineages following a duplication event early in eudicot evolution. In the vast majority of taxa, one or the other form of RPB2 has been lost from the genome. This pattern of duplication and loss, which the data show applies to other nuclear genes as well, may well be responsible for the rapid diversification of flowering plants that occurred during the Paleocene and Eocene.

At a much earlier time, prior to the existence of land plants, the ancestral plant genes for RNA polymerase II protein subunits underwent duplication of a very different sort to encode the enzyme we know as RNA Polymerase IV. This thesis presents comparative sequencing studies and phylogenetic analyses of the genes for the largest Pol IV subunit (RPD1 and RPE1) and for the second-largest Pol IV subunit (RPD2). The results clearly show that both of these Pol IV subunit genes were derived from the corresponding Pol II homologs. These duplication events were separated in evolutionary time, with RPD1 appearing in charalian algae and RPD2 first appearing in primitive land plants.

Extensive purification was carried out with nuclear RNA Polymerase preparations from cauliflower, toward the goal of identifying a catalytically active RNA Polymerase IV. Owing to the low level of this protein, no activity could be demonstrated. Nonetheless, by using immunochemical techniques and mass spectrometric analysis, it was shown that RPD1/RPE1 and RPD2 proteins are closely associated and that the RPB3 gene product, together with common subunits are part of the Pol IV apoenzyme. Pol IV is non-essential for normal growth and development of Arabidopsis. However, Pol IV knockout mutants are defective in asymmetric DNA methylation.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number page

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to thank my major advisor and mentor Ben Hall. None of my dissertation work could be possible without Ben's full support and advice. Over the years, I have benefited from his broad knowledge, his critical thinking and reasoning, his optimism, patience and persistence in scientific researches, and his open-mindedness for new ideas. I also want to thank Liz van Volkenburgh. Liz has been nice and helpful from day one I arrived here. She is always kind and encouraging. Keiko Torii, Willie Swanson, Milton Gordon, Knut Aagaard, and William Trager have served in my supervisory and/or reading committee.

I also want to thank Yajuan Liu, who has been a best friend and colleague. We have had lots of stimulating and inspiring arguments and discussions. Many of them have contributed to my dissertation work. Nori Yoshikawa sequenced most of the RPB2 genes studied in the Chapter 2. John Stiller has given me many suggestions on the Pol IV evolution in the Chapter 3. The mass spectrometry in Chapter 4 was done by Claire Delahunty and John Yates in Scripps Research Institute.

I want to thank my wife, Luzhang Shao, for her love, understanding, and encouragement. My father, Wenyuan Luo, and mother, Zhongrong Liu, have been supportive all the years.

Chapter 1. RNA Polymerase: Structure, Function and Evolution

DNA dependent RNA polymerases (RNAPs), enzymes that use DNA as the template for complementary RNA synthesis, are directly responsible for functional expression of genetic information stored in DNA. RNAPs and transcription processes are the primary targets for gene regulation by environmental and developmental cues. Consequently, transcription processes and their regulators are subject to strong natural selection. We are interested in the function and evolution of RNA polymerases of various types. In this dissertation, I will discuss the application of the sequences of RNA polymerase II genes (RPB1 and RPB2) to the study of angiosperm evolution (Chapter 2), the discovery of RNA polymerase IV in charalean green algae and land plants (Chapter 3), and the characterization of RNA polymerase IV in cauliflower inflorescence (Chapter 4). In this section (Chapter 1), I will review the structure and function of various RNA polymerases, emphasizing the mechanism of RNA polymerase action and the relationships between RNA polymerase polypeptide sequences and the functions they carry out. Then I will discuss evolution rates of RNA polymerase genes and the application of RNA polymerase genes to phylogenetic studies.

1.1 Composition of RNA polymerases

There are two categories of RNA polymerases: the single subunit T7 or T7-like RNA polymerases and the multiple subunit RNA polymerases. The single subunit T7-like RNAPs are found in the genomes of bacteriophages T7, T3, SP6 and K11 (Dietz et al., 1990) and in eukaryotic nuclear genomes. Nuclear encoded T7-like single subunit RNAPs are responsible for transcription of mitochondria genes and some plastid genes (Hedtke et al., 1997; Hess and Borner, 1999). The nuclear encoded T7-like RNAPs have N-terminal transit peptides targeting the RNAP proteins to mitochondria or chloroplasts (Hedtke et al., 1997). There are three T7 like RNAPs in the Arabidopsis genome (RpoT1-3). RpoT1 and T3 are targeted to mitochondria and chloroplasts respectively. RpoT2 is targeted for both mitochondria and chloroplasts, and is the first RNA polymerase shown

to transcribe genes in two different genomes (Hedtke et al., 2000). In angiosperms, the chloroplast rpoT RNA polymerase arose by gene duplication from an ancestral gene encoding the mitochondrial rpoT RNA polymerase (Richter et al., 2002). There are six active T7-like RNAPs in allotetraploid *Nicotiana tobacum* (RpoT1-6). Translation of RpoT3 starts at CUG, a non-ATG translation start codon (Hedtke et al., 2002). The translational product of T7 RNAP is about 100 kD, with conserved domains A through C (Delarue et al., 1990; Sousa et al., 1993).

Cellular multiple subunit RNAPs, containing 5 to 15 subunits, account for most cellular transcription. Eubacterial RNAP core enzyme and plastid encoded bacterial-like RNA polymerases have a largest subunit (beta' subunit encoded by rpoC) and a second-largest subunit (beta subunit encoded by rpoB) and two alpha subunits (rpoA) (Fleischmann et al., 1995). The RNAP genes are organized in a conserved alpha-beta-beta' operon. Beta and beta' subunits are fused into a single polypeptide in *Helicobacter pylori* (Tomb et al., 1997). The beta' subunit in the plant plastid genome is split into two polypeptides (rpoC1 and rpoC2) between the conserved regions E and F (Serino and Maliga, 1998) .

There are three DNA dependent RNA polymerases (RNAPs) in the nuclei of all eukaryotes. Three DNA dependent RNA polymerases have been purified from animals (Roeder and Rutter, 1969; Greenleaf and Bautz, 1975; Sklar et al., 1976), yeasts (Adman et al., 1972; Young and Whiteley, 1975), plants (Jendrisak and Burgess, 1975; Guilfoyle et al., 1976), and protists (Pong and Loomis, 1973). Pol I transcribes rRNA; pol II transcribes mRNA and most SN RNAs; and pol III transcribes tRNA, 5s rRNA and some SN RNAs. Each polymerase has 10 to 12 subunits, with its own unique largest subunit (beta' homolog) (namely RPA1, RPB1 and RPC1 for pol I, II, and III respectively) and second largest subunit (beta homolog) (RPA2, RPB2 and RPC2 for pol I, II, and III respectively), as well as some unique small subunits and shared subunits (Young, 1991; Sakurai et al., 1996). Most small subunits in yeast including RPB6, RPB8, RPB9, RPB10, RPB11 and RPB12 can be substituted by their human homologs (McKune et al.,

1995; Shpakovski et al., 1995). In plants, there is an additional putative polymerase, pol IV, which is one of my dissertation projects. Archaebacteria have one RNAP with a similar subunit composition and extensive homology to eukaryote nuclear RNAPs (Langer et al., 1995; Bell and Jackson, 1998).

The largest and second largest subunits of all multiple subunit polymerases are conserved throughout evolution (Mooney and Landick, 1999; Cramer et al., 2000; Korzheva et al., 2000). There are 8 conserved regions in all the largest subunits, named regions A through H (Jokerst et al., 1989; Puhler et al., 1989). By comparing yeast pol II and *E.coli* RNA polymerase, Sweetser *et al.* identified 9 conserved regions in the second largest subunits, named conserved regions A through I (Sweetser et al., 1987). An additional three conserved domains for second largest subunits have been identified by considering all multisubunit RNA polymerases, and together with domains A through I numbered 1 to 12 (Denton et al., 1998).

1.2 RNA polymerase structure and function

Unlike DNA polymerases (DNAPs), which extend a preexisting primer strand, RNAPs initiate a transcript *de novo* by binding to specific promoter DNA sequences, melting the two strands of the DNA in the vicinity of the start sites, and joining the first two monoribonucleotide residues. The RNAP pre-initiation complex (PIC) is unstable and is characterized by abortive cycling, a process of repeated synthesis and release of 3-8 short nucleotide RNA products. When the RNA-DNA hybrid reaches 8-9 base pairs (bp), the RNAP clears the promoter and undergoes a transition to form a stable, highly processive transcription elongation complex (TEC). One distinguishing feature of transcription (as opposed to DNA replication by DNAPs and RNA replication by RNA dependent RNA polymerases (RdRPs)) is that the RNAP must release the nascent transcript from the template strand. During elongation, the two strands of DNA are separated downstream of the transcription complex and re-annealed at the upstream end, forming a transcription "bubble" that encloses an 8-9 bp RNA/DNA hybrid.

1.2.1 T7 RNA polymerase structure

Single subunit T7 RNAP has an overall architecture, active center structure and mechanism for nucleic acid polymerization similar to those of other single subunit polymerases such as DNA polymerase (DNAPs), RNA dependent RNA polymerase (RdRPs) and HIV reverse transcriptase (RT) (Sousa et al., 1993; Steitz et al., 1993; Temiakov et al., 2004; Yin and Steitz, 2004). The T7 RNAP structure has a deep cleft that resembles a right hand, with palm, thumb and finger domains. Two Aspartate residues (D537 in conserved motif A and D812 in conserved motif C) bind to two metal ions at the active center in the palm domain. The O helix in motif B of the finger domain is involved in binding and discriminating the coming NTP substrates (Temiakov et al., 2004; Yin and Steitz, 2004). As with all other RNAPs, the polymerization cycle consists of NTP binding, polymerization or phosphodiester bond formation, pyrophosphate (PPi) release, and translocation. During each nucleotide addition cycle, T7 RNA polymerase undergoes a transition from a catalytically inactive "open" to an active "closed" conformation by rotation of a tyrosine (Y639) at the end of the O-helix domain. Unlike DNA polymerase whose substrate selection happened in the "closed" state, substrate selection occurs before T7 RNA polymerase undergoes a conformational change. In the "open" state, the substrate binds specifically to the T7 RNAP by pairing to the template base. Y639 interacts with the 2' OH of the incoming NTP via a Mg^{++}-bridge, thus discriminating NTPs from dNTPs (Temiakov et al., 2004). After NTP binding, the O-helix undergoes a conformational change from the "open" to the "closed" state, pushing the NTP close to the active site, where a phosphoryl transfer reaction occurs to produce PPi. The PP_i is bound to a Mg^{2+} ion, and crosslinks the active center D537 to O-Helix; thereby PPi maintains the RNAP in an identical conformation as in the substrate complex. PPi release breaks the crosslinks between the active center and O-Helix and drives RNAP translocation and template strand separation (Yin and Steitz, 2004). This process is called the powerstroke mechanism.

T7 RNA polymerase binds to the promoter from positions -17 to +6 relative to the transcription start site and can form the initiation complex without any accessory factors. The polymerase active site lies near the C-terminus of T7 RNAP. The 324 residues in the N-terminal domain of T7 RNAP, unique to all phage-like RNA polymerases, play a critical role in promoter recognition and DNA melting. This region includes a protein domain that recognizes AT rich DNA around the -17 region, a specificity loop that recognizes sequences around -9, and an intercalation loop that facilitates DNA melting. The active site of the initiation complex can allow only 3 bps of RNA-DNA hybrid (Cheetham et al., 1999; Cheetham and Steitz, 1999). Upon transition from initiation complex to elongation complex, the N-terminal region undergoes a remarkable rearrangement to create DNA-RNA hybrid binding sites which can accommodate 8 bps of hybrid, an RNA exit path, and a more accessible substrate entry pore. The rearrangement involves alternative refolding of 130 residues and reorientation of a stable core subdomain, resulting in formation of three structural elements: the N-terminal extension, a flap-like subdomain and a C-terminal linker connecting the N-terminal domains to the C-terminal enzyme active sites. The N-terminal extension adopts a compact loop-helix-loop-helix conformation that contains sites for RNA-DNA hybrid binding. The flap domain spans the interval between the upstream end of RNA/DNA hybrid and downstream DNA (Tahirov et al., 2002). Thus the N-terminal domain of T7 RNAP plays dual and distinct roles in initiation complex and elongation complex.

1.2.2 Taq RNA polymerase and Yeast pol II structure

Crystal structures of RNAP from *Thermus aquaticus*(Zhang et al., 1999), *Thermus thermophilus* (Vassylyev et al., 2002) and *S. cerereisae* pol II have been resolved (Cramer et al., 2000; Cramer et al., 2001; Armache et al., 2005) (Fig. 1.1). The multiple subunit RNA polymerases have no apparent sequence or structural homology to single subunit RNA polymerases. The overall structures of Taq RNAP and yeast pol II are similar, with the largest and second largest subunits forming the large channel (cleft) that accommodates template DNA and RNA product. Yeast pol II and bacterial RNAP share

the same core structure but differ entirely in peripheral and surface structure, suggesting that they have a conserved catalytic mechanism but different interactions with other proteins such as general transcription factors and regulatory factors.

Yeast pol II has an overall negatively charged outer surface with a uniformly positively charged cleft including the active center, the wall behind the active center blocking the extension of DNA/RNA path and a "saddle" between the clamp and the wall. The structure of yeast pol II can be divided into four mobile modules: the "Core" module, the "Jaw-Lobe" module, the "Shelf" module and the "Clamp" module. The Core module is the major part of the enzyme: containing regions C, D, E, F of RPB1, regions E to I of RPB2 that form the active center, and all of RPB3, RPB10, RPB11 and RPB12. The Jaw-Lobe module is made up of the upper jaw of RPB9 and RPB1 between region G and H, and the lobe region of RPB2 between regions B and C. The shelf module contains RPB5, RPB6 and the "foot" region (between motifs F and G) and "cleft" regions (motifs F, G and H) of RPB1. The clamp module is made up of the C-terminus of RPB2 and the N-terminal and conserved H regions of RPB1. DNA enters the enzyme's cleft between the cleft and jaw-lobe modules. The clamp module holds the DNA and RNA in place. The core module holds the active center in the floor of the cleft. A hole in the cleft underneath the active center ("pore1") allows substrate NTPs to enter and RNA backtracking to occur (Cramer et al., 2000; Cramer et al., 2001). A third of the total surface area buried in subunit interfaces is accounted for by RPB1 and RPB2 contacts, thus accounting for the high stability of pol II. RPB3, RPB8, RPB10, RPB11 and RPB12, mutually interacting with one another and with RPB1 and RPB2, are important for enzyme assembly. The conserved regions of the RPB1 and RPB2 proteins include parts of the enzyme's active center and those sites interacting with each other and with common subunits that are important for enzyme assembly (Cramer et al., 2001).

1.2.2.1 Active center

The template DNA base at the active center is designated as the i+1 position and the downstream nucleotides are designated as i+2, i+3 i+4 and so on. The last previously added nucleotide is designated as i-1 site and the upstream sequences are designated as i-2, i-3, i-4 and so on. The growing RNA 3' end is positioned at the active site, above a pore through which the nucleotides may enter. After the i-1 nucleotide has been added to the RNA, the enzyme undergoes translocation, advancing the i+1 nucleotide and moving the i+2 base of the downstream template strand into the active center. The nucleotide addition site (designated "A" site) opposes the DNA base at site i+1 (Fig 1.2). Compared to the path of the incoming DNA template, nucleotide i+1 is twisted leftward by 90°. Therefore the nucleotide at i+1 points downward toward the floor of the cleft for readout at the active center, while the base at i+2 is directed upward into the opening of the cleft. As a result of the twist, the hybrid helix is at an angle of almost 90° to the incoming DNA helix (Gnatt et al., 2001)(Fig 1.3B).

There are two metal ions at the active center, chelated by the three Aspartate residues in the invariant region D motif (NA**D**F**D**G**D**) of the largest subunit (Cramer et al., 2001; Vassylyev et al., 2002). In yeast pol II, the second metal ion is also bound by the Glutamate and Aspartate residues of the conserved motif GYNQ**ED** in region F of second largest subunit (Cramer et al., 2001; Sosunova et al., 2003). The two metal ions are a common feature among all other polymerases (Steitz et al., 1994; Sosunova et al., 2003). In yeast pol II, the first metal ion (termed "metal A") is tightly bound at the active center. The second metal ion (termed "metal B") has a low level of occupancy in the crystal structure and is recruited together with the substrate NTP (Cramer et al., 2001; Westover et al., 2004).

A conserved α-helix in the motif F (F-bridge) has been observed across the channel near the active center in both eubacterial RNAPs and yeast pol II. In the eubacterial RNAP, the F-bridge is bent or flipped out in the middle, placing its amino

acid side chains in contact with the template DNA at position i+2 (Zhang et al., 1999; Vassylyev et al., 2002). In the yeast pol II structure, the F-bridge is essentially straight both in free enzyme and in transcription elongation complexes (Cramer et al., 2001; Gnatt et al., 2001; Westover et al., 2004). It is proposed by several independent groups that the F-bridge may undergo conformational changes from straight to bent during RNAP translocation, and that the transition between the straight and bent states is crucial for coupling the enzyme conformational change to RNAP translocation (Gnatt et al., 2001; Epshtein et al., 2002; Vassylyev et al., 2002). Protein-DNA cross-linking studies of bacterial RNAP showed that both straight and bent states exist for the F-bridge. This led Goldfarb's group to propose a "Swing-Gate" model in which F-bridge and G-loop act cooperatively in the conformational transition of the F-bridge during nucleotide addition and RNAP translocation (Epshtein et al., 2002). Others have argued against this model, because there is no direct evidence that the F-bridge undergoes a conformational change during chain elongation either in pol II or Taq RNAP. Furthermore, the bending observed in bacterial RNA polymerases would move the invariant residues Thr831 and Ala832 in yeast pol II further from the A site (Westover et al., 2004).

1.2.2.2 Nucleotide addition

Before entering the active center, the downstream double-stranded DNA is unwound into template and non-template strands. The extent of template DNA melting can vary from +3 to +6 residues in different structures (Kettenberger et al., 2004; Westover et al., 2004). These differences may reflect the active vs inactive state of the enzyme. The unwinding of the downstream DNA is facilitated by the positively charged residues in switch 2 (conserved region C in RPB1), which pull the template strand away from the duplex axis, and the negative charged residues in switch 1 (conserved region H in RPB1), which repel the DNA strand. In addition, fork loop 2 (near conserved region D in the RPB2) blocks duplex binding and prevents re-association of the separated strands (Kettenberger et al., 2004). Unwound template DNA is stabilized by the F-bridge helix. The invariant residues Thr831 and Ala832 in the middle of the F-bridge contact the

residue at the i+1 site and residues Tyr836 and Arg839 contact the residue at the i+2 and +3 sites (Gnatt et al., 2001; Kettenberger et al., 2004) (Fig 1.2, 1.3B).

Nucleotide selection/addition in the yeast pol II may occur by a two-step mechanism with initial binding to an entry site (the "E" site) site beneath the active center in an inverted orientation, followed by rotation into the nucleotide addition (the "A" site) (Westover et al., 2004). At the A site, the substrate NTP pairs with the template base at i+1 site. The substrate is coordinated by the two metals at the active site and stabilized by invariant residues surrounding the NTP binding site (Kettenberger et al., 2004; Westover et al., 2004) (Fig 1.3A). The presence of an E site was revealed by binding of a mismatched nucleotide. The orientation of the mismatched nucleotide is flipped, with β and γ phosphates coordinating metal ion B and the sugar and base projecting downwards into the pore beneath the active center (Westover et al., 2004). The residues that interact with the phosphate and sugar overlap with the residues that interact with the nucleotide in the A site. The existence of an E site was also predicted by the exonuclease activity of pol II, which can be stimulated by a mismatched nucleotide (Sosunova et al., 2003). Westover *et al.*(2004) proposed that the NTPs may bind at the E site first and then rotate around the metal B to the A site, where base pairing occurs. There is no direct evidence regarding the mechanism by which RNAP distinguishes NTP from dNTP. Cramer and his colleagues proposed that invariant residue N479 at the active center may interact with the 2'OH to distinguish the NTPs from dNTPs (Cramer et al., 2001; Kettenberger et al., 2004), however other groups have noted that N479 is too far away from the 2'OH for hydrogen bonding (Westover et al., 2004). Westover et al. (2004) proposed that binding to the E site and nucleotide rotation may play a role in NTP/dNTP discrimination, but supporting evidence for this is also lacking.

1.2.2.3 DNA-RNA hybrid

The RNAP channel can hold nine base pairs of DNA-RNA hybrid between the F-bridge and the "wall" (between conserved regions F and H of RPB2). The first base pair

of hybrid is at the i+1 site. The template DNA is bound by protein over the entire length of the hybrid. The first three ribonucleotides around the active center are contacted by protein and are rigidly held, to insure fidelity of transcription. The five upstream ribonucleotides are held mainly by base pairing. All protein-DNA and protein-RNA interactions involve the sugar-phosphate backbone rather than specific base interactions. There are approximately 20 positively charged side chains around the hybrid. These form a shell, attracting the hybrid but not contacting or restraining it (Gnatt et al., 2001).

Fifteen protein regions are involved in contacts with the DNA/RNA hybrid. The 3' end of the hybrid is bound by the active site in RPB1 and "switch 3" in the conserved region I of RPB2. The hybrid binding region in conserved region H and I of RPB2 contacts the newly transcribed RNA; the hybrid binding region in the conserved region E and F of RPB2 binds the template DNA of the hybrid (Kettenberger et al., 2004) (Fig 1.2, 1.3B).

1.2.2.4 The RNA exit channel

The DNA-RNA hybrid is eight base pairs in length and the paths of the RNA and DNA begin to diverge at position -8, with residues -9 and -10 completely separated from the complementary DNA (Kettenberger et al., 2004; Westover et al., 2004). Three protein loops, the "lid', the "rudder" and the "fork loop 1", play key roles in RNA-DNA separation. The lid (conserved region B in RPB1) serves as physical barrier driving the DNA and RNA strands apart, maintaining the separation of the strands and guiding the RNA along an exit path (Fig 1.3B). Residue Phe252 in RPB1 splits the RNA-DNA base pair at position -10, contacting the DNA base with the plane of the aromatic side chain perpendicular to the plane of the base. RPB1 residue Phe264 may similarly contact the DNA base at position -10 or -11. The lid sequences are not conserved between different RNA polymerases. The corresponding residue of Phe252 in polymerases other than pol II is a hydrophobic residue (Leucine or Methionine). The rudder (proceeding conserved region C in RPB1) interacts with the DNA at positions -9, -10 and -11, preventing re-

association with the RNA. The fork loop 1 (before conserved region D in RPB2) interacts with RNA at position -5, -6 and -7 in the hybrid region, preventing unwinding of hybrid past position -8. RNA exits the active center regions through an exit tunnel formed by the lid and the saddle between the wall and the clamp (Kettenberger et al., 2004; Westover et al., 2004).

1.3 The evolution rate of RNA polymerase genes

It has been known for decades that the evolutionary rates of different proteins vary over several order of magnitude, but the reasons for this are not fully understood. The neutral theory of molecular evolution predicts that the rate of amino acid or nucleotide substitution is approximately constant per site per year and the rate of evolution ('conservative nature' of changes) depends on the importance of the proteins and the nature of the mutations. The rate of evolution should be greater in proteins that contribute less to individual fitness and in sites under weaker purifying selection (Kimura and Ohta, 1974; Kimura, 1991). If protein evolution is in part due to slightly deleterious amino acid substitution, then the rate of evolution should depend on the dispensability of the entire protein to the organism. In proteins that make a smaller contribution to the organismal fitness, a large proportion of mutations would fall within the range that could be considered neutral. Therefore, if some molecular evolutionary change is caused by genetic drift, and the purifying selection is reduced, mildly deleterious substitutions would accumulate more rapidly and the rate of evolution should be higher in proteins that are less important and more dispensable to the organism (Kimura and Ohta, 1974; Wilson et al., 1977; Kimura, 1991). Based on the same rationale of the dependence of evolutionary rate on protein dispensability, it is argued that the rate of protein evolution is also dependent on the proportion of potential amino acid changes that are compatible with proper protein function (Wilson et al., 1977). Any change that perturbs the activity of the protein must be selected against, or subsequently compensated for by a correlated change in the interacting partner. Thus, the higher the proportion of sites that are constrained by function or the more protein-protein interactions that a protein has with

interacting partners, the less is the chance that a mutation will be compatible with function and the slower the evolutionary rate will be. However, these predictions have been hard to test because it is difficult to measure the dispensability and interactions of a protein and it is difficult to evaluate the difference statistically.

Recent work using high throughput databases confirms that protein dispensability, functional density, expression and modularity (protein-protein interactions) all significantly contribute to the rate of evolution (Herbeck and Wall, 2005), but the relative importance of each factor is unknown and may vary from case to case. To estimate protein dispensability, Hirsh and Fraser (2001) used growth rates of yeast strains in which individual genes were deleted. From the 548 mutants that have fitness effect ranging from 0 to 0.5 (0 represents no effect, 1 represents lethal), they found a highly significant relationship between protein dispensability and evolutionary rate, estimated from evolutionary distance between *S. cerevisiae* and *C. elegans* (Hirsh and Fraser, 2001). Comparing the synonymous (Ks) and nonsysnonymous (Ka) substitution rates and Ka/Ks ratio in bacterial pairs of essential and nonessential genes, Jordan et al., 2002 found that the average Ka, Ks and Ka/Ks are significant higher in nonessential genes than essential genes (Jordan et al., 2002). Wall et al. 2005 used >3,000 proteins in four species of yeast genus *Saccharomyces* whose genomes have been sequenced to investigate the relationships between the level of expression and protein dispensability. They found that both dispensability and expression have significant but independent effects on the rate of protein evolution (Wall et al., 2005).

Experimental advances including development of the yeast two-hybrid method and mass spectrometry make it possible to characterize protein-protein interaction networks in cells. Using a list of 3541 interactions between 2445 different yeast proteins and comparing the evolutionary distances of orthologs between *S. cerevisiae* and *C. elegans*, Fraser et al, 2002 found a significant negative correlation of the protein interactions (measured by the number of interactions) with the rate of evolution (measured by the evolutionary distance of orthologs) (Fraser et al., 2002). Later, this group used a more

complete interaction dataset to make similar comparisons (rate of evolution vs number of interactions) using data from *S. cerevisiae*, *Candida albicans* and *Schizoscaccharomyces pombe*. This study confirmed that the number of interactions correlates negatively with the rate of evolution (Fraser et al., 2003). At sites important for interaction between proteins, evolutionary changes may occur largely by co-evolution, in which substitutions in one protein result in selective pressure for reciprocal changes in interacting partners. As a consequence of co-evolution, mutually interacting proteins evolve at similar rates (Fraser et al., 2002). However, not all interactions may lead to co-evolution. Residues in the interfaces of obligate complexes tend to evolve at a relatively slower rate, allowing them to co-evolve with their interacting partners. In contrast, transient interactions have a plastic evolutionary rate. There is little or no evidence of correlated mutations for transient interactions (Mintseris and Weng, 2005).

There are unique largest and second-largest subunits for each eukaryotic RNA polymerase. After evolution and functional divergence of the three RNA polymerases in early eukaryotes, the genes for the largest subunits evolved independently of one another for RNAP I, II and III and likewise for the second-largest subunits. Each RNA polymerase has very distinct protein-protein interactions with its transcription factors and different mechanisms for regulating gene transcription. One of the intriguing questions in RNA polymerase evolution is whether there are differences in the evolutionary rates of the genes for the largest and second largest subunits of RNAP I, II and III. To study the evolutionary rate of RNA polymerase genes in different lineages, we compared the pair-wise substitution rates of protein sequences of the largest and second largest subunits of pol I, II and III in fungi and animals (Fig. 1.4). The substitution rates for the largest and second largest subunits of pol II are, normally but not in all cases, lower than those of pol I and III. To avoid using combinations of species that were either too closely related or too divergent, we used the pair-wise substitution rate in the range of 0.05 to 0.5 substitution per site. To compare the rate differences, we calculated the ratio of substitution rates of the largest and the second largest subunits of pol I and III against those of pol II in the same pairs of organisms (Fig 1.4).

The rates of evolution of RNA polymerase genes are different in fungi and in animals. The substitution rates of the largest and second largest subunits of Pol I and III (RPA1, 2 and RPC1, 2) are similar to those of pol II (RPB1, 2) in fungi. The average ratios of substitution rates are 1.41±0.39 for RPA1/RPB1, 1.09±0.31 for RPC1/RPB1, 1.52±0.35 for RPA2/RPB2, and 1.13±0.3 for RPC2/RPB2. Similar evolution rates of the largest and second largest subunits of Pol I, II and III suggest that functional divergence and regulatory differences between the three RNA polymerase systems have only a small impact on the rate of evolution of the genes for the largest and second largest subunits of the three RNA polymerases in fungi. In other words, the different functional constraints of Pol I, II and III do not greatly affect the evolution rates of RNA polymerase genes in fungi. However, the situation is greatly different for the evolution rates of RNA polymerase genes in animals. The average substitution rate of RPA1 is 4 times higher than that of RPB1 (3.94 ±1.87) and the average substitution rate of RPA2 is 7-8 times higher than that of RPB2 (7.81±3.43). The average substitution rates of RPC1 and RPC2 are slightly higher than those of RPB1 and RPB2 in animals (1.83±0.29 for RPC1/RPB1, 2.87±1.56 for RPC2/RPB2). Similar comparisons for the RNA polymerase genes of Arabidopsis and Rice also shows difference in the substitution rates of genes for pol I, III and IV as compared to pol II (Chapter 3). So how can we explain different substitution rates of RNA polymerase genes in different RNA polymerases and in different lineages?

Since all three RNA polymerases transcribe parts of the genome that are necessary for the survival of cells, essentiality should play a similar role for the evolution rates of genes in all three RNA polymerases. The large number of protein-protein interactions during RNA polymerase assembly and during various stages of the transcription reaction may contribute to the general conservation of RNA polymerase genes across different taxa. In general, more protein-protein interactions and more regulatory effects are observed in pol II transcription, which may explain the reasons for comparatively slower evolution of RNAP II genes. However, because RNA polymerase genes in fungi have smaller differences in the relative evolution rates of Pol I, II and III, protein-protein

interactions alone cannot fully explain the difference in evolution rates of RNA polymerase genes in animals and plants. Compared to the yeast genome, the genomes of animals and plant are larger and contain more genes. More regulatory interactions are expected for Pol II transcription in animals and plants. Additional functions for pol II, such as transcription of microRNA in plants and animals, might also affect the evolution rate of pol II genes.

1.4 The Application of RPB1 and RPB2 genes used in phylogenetic studies

RNA polymerase genes, especially those encoding the largest and second largest subunits of pol II (RPB1 and RPB2), are good candidate genes for phylogenetic studies. First, both the largest and second largest subunits are large proteins with many informative sites. The largest subunit genes contain ~ 1500 AA residues and the second largest subunit genes contain ~ 1200 AA residues. Both conserved regions and non-conserved regions can be used for phylogenetic analyses at various taxonomic levels. Secondly, RPB1 and RPB2 genes are slowly-evolving genes. There is no obvious rate heterogeneity between different lineages. Third, these genes occur either as single copy or as closely related paralogs in most organisms, including animals, fungi, red algae, protists, and most plants. Up to now, there have been no reports of RNA polymerase II isoenzymes or differential usage of the largest or second largest subunits of RNA polymerases. The largest subunit and second largest subunit genes of RNA polymerase II (RPB1 and RPB2) have been used in phylogeny of early eukaryotes, fungi and arthropods and some green plants (Stiller and Hall, 1997; Denton et al., 1998; Stiller et al., 1998; Liu et al., 1999; Shultz and Regier, 2000; Dacks et al., 2002; Liu and Hall, 2004; Nickerson and Drouin, 2004). In this dissertation, I will report the only major exception to the rule of single-copied RPB1 and RPB2; these occur in some angiosperms. We have found paralogs of RPB1 and RPB2 genes in some core eudicots and differential usage of RPB2 genes in some asterids (Chapter 2). We think paralogy might be a common feature for most of the nuclear genes in angiosperms, due to the prevalence of polyploidization and gene duplication events in plants.

RPB1 and RPB2 gene sequences are most successfully used in phylogeny at the sub-kingdom level. Although pol II functions to transcribe pre-mRNA and SnRNA in all eukaryotes, different genome organization and differences in pol II regulation can profoundly influence the evolution of RPB1 and RPB2 genes in different kingdoms. At a sub-kingdom level, genome organization and regulation are similar, slowing the evolutionary divergence of RPB1 and RPB2 genes. In general, caution should be exercised in interpreting RPB1 and RPB2 gene trees for inter-relationships that span different kingdoms. This cautionary rule may apply as well to other genes for studying relationships between different kingdoms.

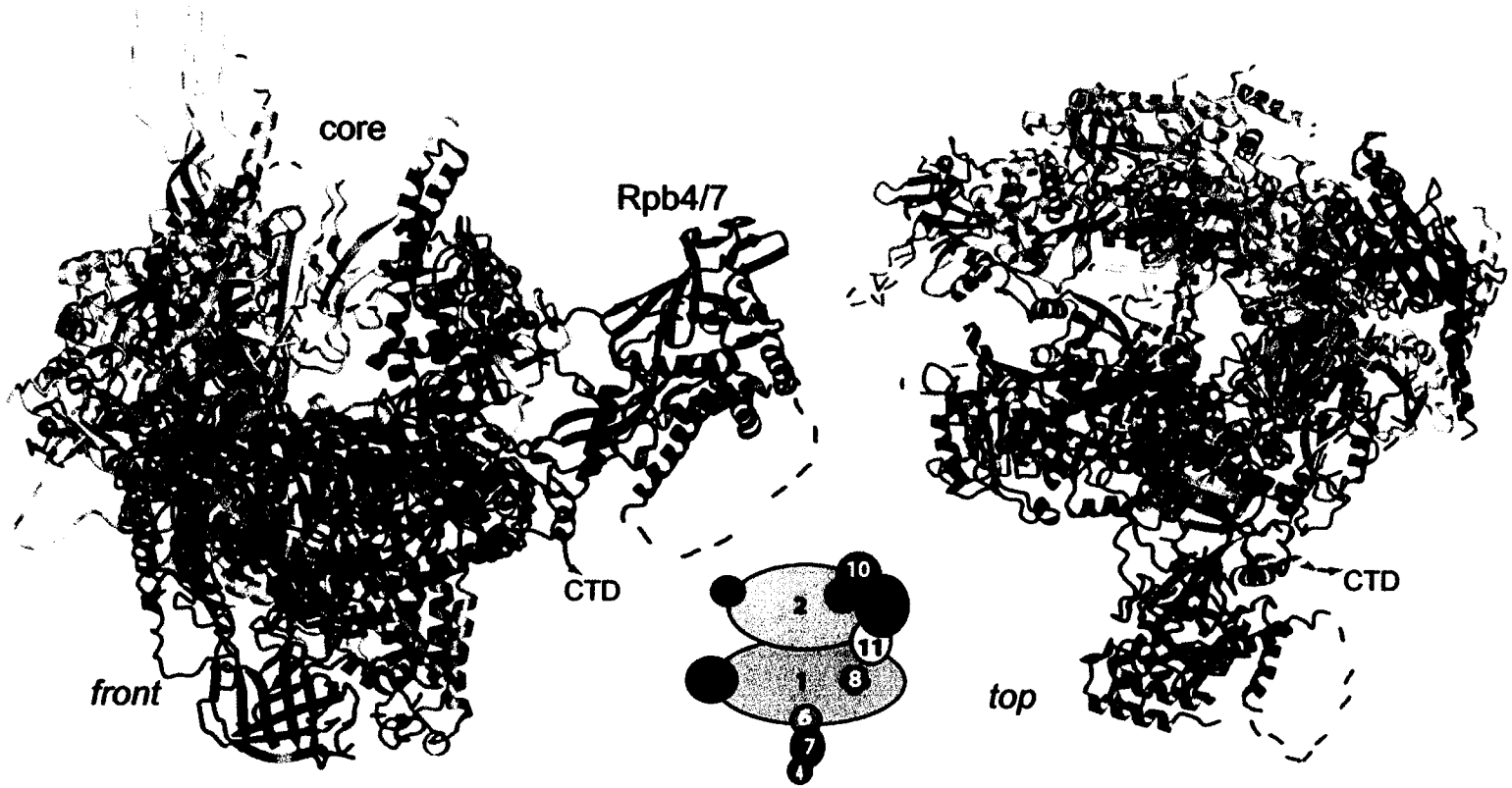


Figure 1.1 The complete *S. cerevisiae* pol II structure from Armache et al., 2005. The 12 subunits RPB1-RPB12 are colored as the diagram below.

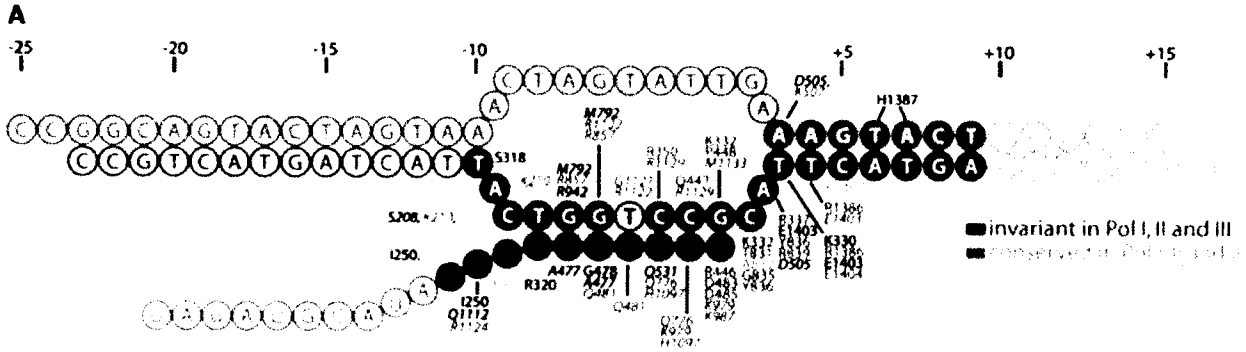


Figure 1.2 Nucleic acids in the pol II elongation complex and pol II interactions. The template DNA strand, nontemplate DNA strand and product RNA are colored blue, cyan, and red, respectively. Filled and open circles denote nucleotides that are ordered and disordered, respectively. Pol II residues that are within 4Å distance of nucleic acid are depicted. Residues that are invariant, conserved, and differing among the three yeast nuclear RNA polymerases are in dark green, light green, and black respectively. Italics are used to distinguish RPB2 from RPB1 residues. This figure and figure legend are from Kettenberger et al., 2004.

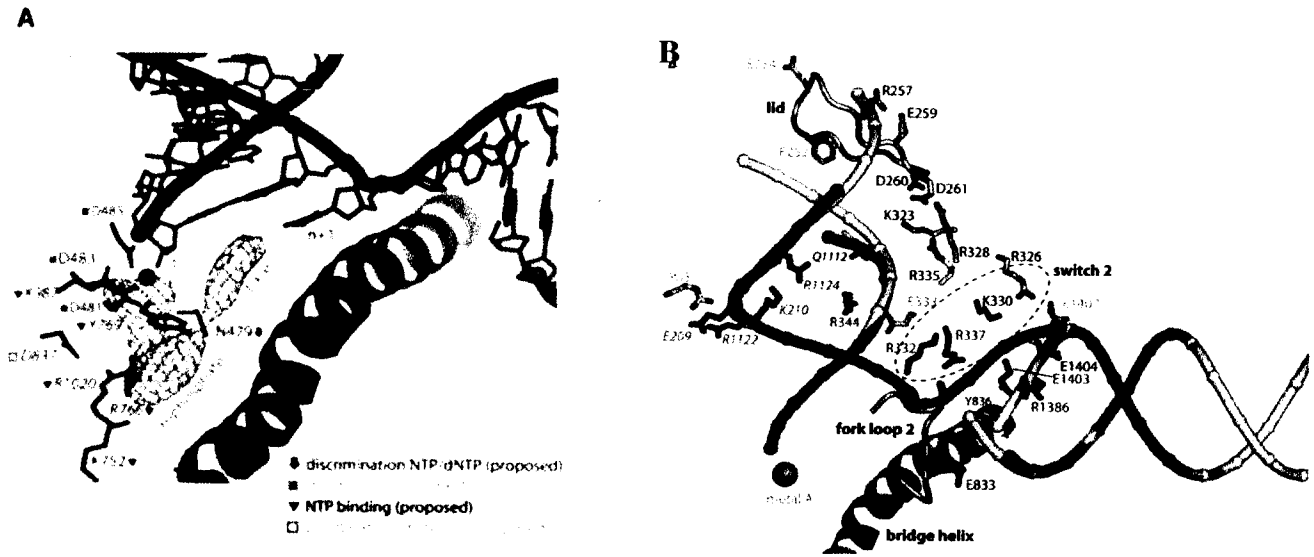


Figure 1.3 (A). NTP binding sites. Invariant pol II residues in the vicinity of the electro density for a GTP substrate analog are shown. (B). DNA unwinding and DNA-RNA separation. Downstream DNA as a canonical B form duplex is in light magenta. The active site metal ion A is shown as a pink sphere. Pol II residues that are apparently involved in nucleic acid separation are colored according to their conservation in pol I, II, and III as in Figure 1.2. The bridge helix, fork loop 2 and the lid are depicted. Basic amino acid in switch 2 that may pull the template strand upwards are encircled with a dashed line. These figures and figure legends are from Kettenberger et al., 2004.

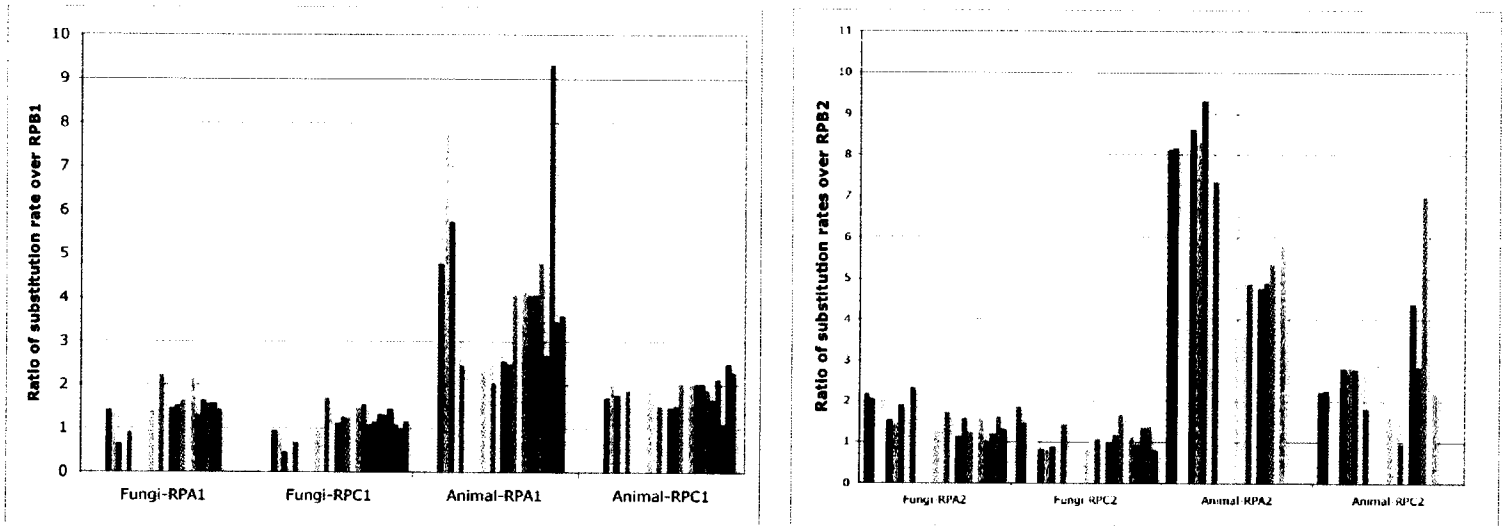


Figure 1.4 The ratios of substitution rates of the largest and second largest subunits of RNAP I and III over those of RNAP II in fungi and in animals.

Chapter 2. Duplication and Paralog Sorting of *RPB2* and *RPB1* genes in Core Eudicots

Summary

RPB1 and *RPB2*, which encode the largest and second-largest subunits of RNA polymerase II respectively, are essential single copy genes in fungi, animals and most plants. Two paralogs of the *RPB2* gene have been found in some groups of plants (Oxelman et al., 2004). Here we report the continuing effort to identify the origin of the *RPB2* gene duplication, focusing on the taxa of lower eudicots. Through careful sampling and phylogenetic analysis, we are able to construct the *RPB2* gene tree in angiosperms and infer the phylogenetic positions of the gene duplication and gene loss events that occurred. Our study shows that an *RPB2* gene duplication occurred early in core eudicot evolution, at or near the time of the *Buxaceae/ Trochodendraceae* divergence. Subsequently, multiple gene duplication and paralog sorting events happened independently in different core eudicot taxa. Differential expression of two *RPB2* gene paralogs may explain the preservation of both paralogs in the asterids. While they are both expressed, the pattern across plant tissues is different. One gene (*RPB2-i*) accounts for most of the *RPB2* mRNA made in the flower organs while the other gene (*RPB2-d*) is predominantly used in the vegetative tissues. We also found two paralogs of the *RPB1* gene in some core eudicot species. The *RPB1* gene duplication occurred before core eudicot divergence, around the time of *RPB2* gene duplication. Several independent *RPB1* paralog sorting events happened in different core eudicot taxa and their occurrence was independent of the *RPB2* paralog sorting events. Our results suggest that a polyploidization event happened at or near the time of the *Buxaceae/ Trochodendraceae* divergence. We propose that this polyploidization and the partial diploidization processes thereafter may have been the driving force of core eudicot radiation.

Key words: RPB2; RPB1; *eudicots; Gene duplication; Paralog sorting; Differential expression.*

2.1 Introduction

Our understanding of phylogenetic relationships among flowering plants has been greatly improved by molecular phylogenetic analyses. Most studies have employed plastid gene sequences (*rbcL*, *atpB*) and nuclear rDNA (*18s rDNA, 26s rDNA*) sequences to infer angiosperm phylogeny (APG, 1998; APGII, 2003). The use of nuclear protein coding genes for angiosperm phylogeny has been hampered by the prevalence of polyploidization and gene duplication events in plants (Masterson, 1994; Wendel, 2000). The paucity of phylogenetic analyses of nuclear protein coding genes restricted further inference of the nuclear evolutionary history and understanding the mechanistic force for angiosperm evolution. We seek to use nuclear protein coding genes to better understand evolution of the eudicot nuclear genome.

DNA dependent RNA polymerase (RNAP) is the principal enzyme responsible for gene transcription, the first step of gene expression and a major target of developmental and environmental regulation. The largest and second largest subunits of RNA Polymerase II and of other prokaryotic and eukaryotic RNA polymerases contain the conserved motifs that make up this enzyme's catalytic surface (Mooney and Landick, 1999; Korzheva et al., 2000). The multiple interactions of these the two subunits with one another, their interactions with other factors (Ishihama et al., 1998), and the global nature of Pol II function imply coding either by a single gene or by highly constrained duplicate genes (Denton et al., 1998). Accordingly, the largest subunit and second largest subunit genes of RNA polymerase II (*RPB1* and *RPB2*) have been successfully used in phylogenetic study of early eukaryotes, red algae, fungi and arthropods (Stiller and Hall, 1997; Stiller et al., 1998; Liu et al., 1999; Shultz and Regier, 2000; Dacks et al., 2002; Liu and Hall, 2004).

Broad-scale phylogenetic studies with *RPB1* and *RPB2* have also been made in green plants for a limited number of species in bryophytes, ferns, gymnosperms and a few lower angiosperms (Denton et al., 1998; Nickerson and Drouin, 2004). Studies of 35

eudicot taxa show that there are two RPB2 paralogs in Gentianales, Lamiales and Solanales of Asteridae I (Lamiids) and in Ilex (Aquifoliales) of Asteridae II (Campanulids). One paralog (*RPB2-i*) has all 24 introns at the same locations as in *A. thaliana*, while the other paralog (*RPB2-d*) lacks introns 18-23 (Oxelman et al., 2004). In Escallonia (of Escalloniaceae in Campanulids) and, as we report here, in Ericales both paralogs have all the 24 introns. Only *RPB2-d* homologs have been preserved in all other core eudicots, including the rest of campanulids and rosids (Oxelman et al., 2004). This report presents our continuing studies on the phylogenetic position of the *RPB2* gene duplication, focusing on the taxa of lower eudicots. Only a single copy of the *RPB1* gene has been found in lower angiosperm species as *Amborella, Nymphaea* and *Magnolia* (Nickerson and Drouin, 2004). Here we report that two paralogs of *RPB1* genes also exist in some core eudicot species. The *RPB1* and *RPB2* duplication events occurred at nearly the same time, shortly before core eudicot divergence. This, together with other evidence, suggests that ancient polyploidy may have played a role in shaping the eudicot nuclear genome.

2.2 Material and Methods

2.2.1 Taxon sampling

We sampled representatives from all major taxa of lower eudicots (Table 2.1). The *RPB2* PCR primers were described previously (Oxelman et al., 2004). Additional primers are: E7F (exon 7 forward) CTTGCYGGYCCTYTRCTKGGWGG;

E7R (exon 7 reverse) CCWCCNAGYARAGGRCCRGCAAG;

E12R (exon 12 reverse) GGRTDHGCNRCNGAHCCNACNG;

GPE2F (exon 2 forward) CCNATGATGACRGARTCWGATGG. Degenerate *RPB1* primers were described previously (Stiller and Hall, 1997).

2.2.2 DNA and RNA extraction

FastDNA kit (Qbiogene #6540-400) was used for plant DNA extraction. Leaf

samples were homogenized by the FastPrep® FP120 Instrument (Qbiogene). Depending on the tenderness of samples, different combinations of lysing matrix were used to optimize the yield. The FastPrep shaking speed was set at 5.0 for 30 seconds. For RNA extraction, tissue samples were homogenized also by the FastPrep® FP120 Instrument in the FastRNA tubes-Green (Qbiogene #6040-601). One more 1/4" cylinder was added for the tough samples such as roots and stems to better homogenize the samples. The shaking speed was set at 6.0 for 30 seconds. Additional shaking was applied as needed to the tough samples. All the solutions and RNA purification columns were using RNeasy Plant Mini Kit (Qiagene #74904). After RNA extraction, DNA-free kit (Ambion #1906) was used to get rid of DNA contamination by Dnase I digestion followed by inactivation of Dnase I.

2.2.3 Phylogenetic Analyses

Modeltest v3.04 (Posada and Crandall, 1998) was used to estimate the DNA substitution model. GTR+I+Γ model was found to be the best-fit model and was used for maximum likelihood (ML) analysis by PAUP* (Swofford, 2002) and Bayesian inference by MrBayes3.0 (Ronquist and Huelsenbeck, 2003). The gamma distribution was separated into four discrete rate classes. For ML analysis, model parameters were first set to the values estimated by Modeltest v3.04 and then re-optimized by PAUP* after an initial heuristic search. A second heuristic search was done with the optimized parameters to get the final topology for ML tree. The heuristic search was done with 20 replicates with stepwise random addition and TBR branch swapping. The nonparametric bootstrap analysis under ML criterion was done for 100 pseuoreplicates, random addition, TBR branch swapping and optimized parameters. The analysis took 25 days to finish. Bayesian inference was conducted by MrBayes 3.0 using uniform prior probabilities and GTR+I+Γ model. We ran four chains for 1000,000 generations and sampled every 100 generations. The trees were summarized by 50% majority rule consensus after initial 2,000 "burn-in". Parsimony analyses were carried out by PAUP* with equal weight, stepwise random taxon addition, and TBR branch swapping.

To compare alternative hypotheses statistically, the one tailed Shimodaria-Hasegawa nonparametric bootstrap test (Shimodaira and Hasegawa, 1999) and parametric bootstrap test under MP criterion (Stefanovic and Olmstead, 2004) were conducted. Shimodaria-Hasegawa tests were conducted by PAUP* using 1000 bootstrap replicates and full parameter optimization of the model. A set of seven hypotheses were tested against the optimal ML tree (Table 2.2). For each hypothesis, a new heuristic search under constraint was carried out by PAUP* to get the ML tree topology under constaint. We followed the procedure for parametric bootstrap test under MP criterion laid out by Stefanovic and Olmstead, 2004.

2.2.4 RT-PCR

DisplayThermo-RT kits (Display system#570-100) were used for the first strand cDNA synthesis. Oligo dT primer was used for first strand cDNA synthesis. Specific primers were designed to differentiate *RPB2-d* and *RPB2-i* genes in the PCR reaction. *RPB1* has been used as a control for PCR reaction and RNA amount. The different primer sets used for RT-PCR were shown as follows: For *Solanum lycopersicum*, RPB2-d forward TATTGGACGAGAAGGGAAACTGGCC, reverse CCCTCGTCAGGTGATTCCCGTTG; RPB2-I forward ACCAGGGCAGGAGTTTCACAGGTTT, reverse GCCAGAATCTTCAGATGACTCCCTC; RPB1 forward D ACTAGGATATAAGGTTGAGAGGCAC, reverse GAGCCTTATTTAACACCTGGTTCAC. For *Antirrhinum majus*, RPB2-d forward CATAGGGCGTGAAGGAAAATTGGCT, reverse CCATCATCAGGAGATTCTCTTTG; RPB2-i forward GACAGCTGAGGAGACAGGTTGATGT, reverse CATAGGCTAGTGTATCCATTCGT; RPB1 forward CTAGGATATAAGGTTGAGCGGCACT, reverse GCCTTATTCAACACCTGATTCACTC. For *Rhododendron macrophyllum*, RPB2-d

forward TTGGGGTAGTTAGAGATATCCGCTT, reverse GCCGATAACTACATCTTCCCCTGAA; RPB2-i forward AGTTATCCGTGACATTCGTCTGAAA, reverse GCGTGTACCGTGCAGTTTGTCCTTG; RPB1 forward AACAGACCTGTCATGGGTATTGTG, reverse CCTTGTTCAGCACCTGGTTCACTC.

2.2.5 Enzymatic cutting

In order to determine the molar ratio of *RPB2-d/i* expressed in the same organs, primers for RT-PCR were designed to match the same regions between the two copies and one-cut enzymes were chosen to distinguish different lengths. The primers and enzyme cutting are shown below. First, RT-PCRs were done to amplify both copies in the same RNA preparation. Then the products were purified and digested by the one-cut restriction enzyme. The digests were separated by electrophoresis stained by ethidium bromide. The band density was calculated by NIH image (http://rsb.info.nih.gov/nih-image/) . For Solanum lycopersicum, forward primer ATGTCACCAA(C/T)TACCAATTTCGGA, reverse primer TCCTTTGTAAGCATTCTTGG, full length cDNA, 1309bp. Restriction enzyme: EcoRI, RPB2-d digested fragments 1180/129 bp, RPB2-i digested fragments 566/742 bp; Taq1, RPB2-d digested fragments, 709/600bp, RPB2-i digested fragments, 1186/122 bp. For Antirrhinum majus, forward primer TGTCCTGCAGAAAC(A/T)CCTGAAGG, reverse primer TCGAATACCTCAAAT(A/T)GGAGACAA, full length cDNA 1281 bp. Restriction enzyme: Ban I, RPB2-d digested fragments 117/1163 bp. RPB2-i digested fragment: 973/301 bp. For Rhdodendron macrophyllum, forward primer ACTTATAC(C/T)CACTGTGAAATTCA, reverse primer ATCTTGTCATC(G/T)ACCATGTGCTT, full length cDNA 1094 bp. Restriction enzyme: Bbs I, RPB2-d digested fragments 774/320 bp, RPB2-i digested fragments 530/564 bp.

2.2.6 Homology Modelling

The *Saccharomyces cerevisiae* pol II structure is used as template (1i50.pdb) (Cramer et al., 2001). The protein structure was viewed and manipulated by Swiss-Pdbviewer (http://swissmodel.expasy.org/spdbv/). The modeled RPB2 protein sequences were aligned with the *S. cerevisiae* RPB2 sequence by Pdbviewer and the alignment was further adjusted by eye. The alignment was then submitted to SWISS-MODEL (Guex and Peitsch, 1997) (http://swissmodel.expasy.org//SM_OPTIMISE.html) for homology modeling.

2.3 Results

2.3.1 Intron-Exon Structure of RPB2 *genes*

Two *RPB2* genes have been found in the Ericales and Lamiids. The *RPB2-d* genes in the latter group have lost introns 18 to 23 (Oxelman et al., 2004). In *Rhododendron macrophyllum*, we sequenced both the *RPB2-d* and *RPB2-i* genes and confirmed the intron positions by RT-PCR. All 24 introns have been kept in both *Rhododendron RPB2* genes. Partial sequencing of *RPB2* genes in *Camellia* and *Diapensia* also confirmed that the *RPB2-d* genes in these species have introns 18 to 23. Throughout this chapter, only exon sequences were used in phylogenic analyses, because the introns generally could not be aligned between plant familes.

2.3.2 The Phylogeny of RPB2 *genes in Eudicots*

The combined exon sequences from exon 2 to 24 of *RPB2* (the "long" dataset), with 2859 total characters, were used for phylogenetic analysis. Both ML and Bayesian inference resulted in the same tree topology. Fig. 2.1 shows the *RPB2* gene tree inferred by ML. Parsimony analyses (MP) resulted a single tree of 8329 steps (Fig. 2.2A). For this tree, the consistency index (CI) = 0.288 and the retention index (RI)=0.433. As compared

to the tree from ML analysis and Bayesian inference, the MP *RPB2* tree differs in two branches: the Ranunculales diverged earlier than Proteales, and *Gunnera* is sister to the *RPB2-i* gene of asterids. In neither case is there support for the branches involved.

Both ML and MP *RPB2* gene trees indicate the occurrence of an *RPB2* gene duplication early in the evolution of core eudicots. Dicotyledous plants that evolved subsequent to this duplication have *RPB2* genes that fall into one of two clades, labeled *RPB2-d* and *RPB2-i* on Fig 2.1. The *RPB2-i* clade contains the genes of *Dilleniaceae*, *Gunneraceae*, *Aextoxicaceae*, *Berberidopsidaceae*, *Vitaceae* and the *RPB2-i* genes of Ericales and euasteridae I. The *RPB2-d* clade contains the *RPB2* genes of *Trochodendraceae*, Saxifragales, *Caryophyllales*, Rosids and the *RPB2-d* genes of Asterids. Each of these clades has strong statistical support. For the *RPB2-d* clade, the posterior probability (PP) is 1.0, the bootstrap support for ML (ML) is 90% and the bootstrap support for MP (MP) is 72%. For the *RPB2-i* clade, the PP =1.0, ML=98% and MP = 82%.

In order to check whether the grouping of *RPB2-i* clade of *Dilleniaceae*, *Gunneraceae*, *Aextoxicaceae*, *Berberidopsidaceae*, and *Vitaceae* is meaningful or is an artifact created by long branch attraction of the *RPB2-i* genes of asterids, we performed phylogenetic analysis after removal all the *RPB2-i* genes of asterids. Deletion of these genes resulted in a single tree with 6936 steps, CI=0.316, RI=0.418. The major features of the tree topology are not affected by the removal of asterid *RPB2-i* genes (Fig 2.2B). The MP bootstrap is 74% for the D clade and 73% for the rest of I clade. This strongly suggests that the grouping of *RPB2-i* and *d* clades is not due to long-branch attraction by the *RPB2-i* genes of asterids.

Tetracenton RPB2 is at the base of *RPB2-d* clade (PP=1.0, ML=90% and MP=72%). Removal of *Tetracenton RPB2* from the dataset does not affect the tree topology elsewhere. If *Tetracenton RPB2* is not included in the *RPB2-d* clade, support for the *RPB2-d* clade increases. In the *RPB2-d* clade, Saxifragales *RPB2* genes are sister to

the *RPB2* genes of Rosids, Caryophyllales and Asterids (PP=1.0, ML=79% and MP=86%). Bayesian inference suggests that Carophyllid *RPB2* genes are sister to Rosid *RPB2* genes (PP=0.96). In the *RPB2-i* clade, *Dilleniaceae RPB2* comes out at the base of the clade (PP=1.0, ML=82%, MP=51%). *Berberidopsidaceae* and *Aextoxicaceae* are grouped together with strong support. The position of *Vitaceae* and *Gunneraceae RPB2* in the *RPB2-i* clade is strongly supported, but the relationships between them and with the rest of I clade are not well supported.

Two different statistical approaches: the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999; Goldman et al., 2000) under the ML criterion and the parametric bootstrapping test (Goldman et al., 2000) under MP criterion (Stefanovic and Olmstead, 2004), have been applied to evaluate alternative hypotheses for placing the duplication event (Table 2.2). The hypothesis that *Gunnera* and *Vitaceae RPB2* genes are members of *RPB2-d* clade is strongly rejected by both analyses. *Dilleniaceae RPB2* as a member of *RPB2-d* clade is rejected by the parametric bootstrap test but marginally accepted by the SH test. Because of the conservative nature of the SH test (Goldman et al., 2000), *Dilleniaceae RPB2* is effectively rejected from the *RPB2-d* clade. The hypotheses that Tetracenton diverged before the *RPB2* gene duplication or that the *Tetracenton RPB2* gene is a member of the *RPB2-i* clade both are not rejected, suggesting that the *Tetracenton RPB2* gene shares features both with the *RPB2-d* and *RPB2-i* genes. This may imply that the duplication happened around the time *Trochodendraceae* diverged. An alternative possibility is that duplication occurred before divergence of the Proteales, since the presence of Proteales RPB2 genes in the *RPB2-d* clade also is not rejected (Table 2.). We conclude that the *RPB2* gene duplication happened at or prior to the divergence of *Trochodendraceae*.

2.3.3 RPB2 *Gene duplications in Ranunculales*

Multiple copies of *RPB2* genes exist in some species within Ranunculales. In order to determine whether these paralogs stem from the same event as the *RPB2-d* and *-i*

duplicate genes, we have sequenced the *RPB2* gene from exon 11 to exon 17 in 16 species of Ranunculales. Exon sequences from exon 11 to exon 17 of *RPB2* (the "short" dataset), with 666 characters, were used for phylogenetic analysis. The phylogenetic trees from Bayesian inference and from ML have the same topology (Fig 2.3). MP analysis resulted in 100 trees in 4 islands of 4051 steps (CI=0.185, RI=0.456). The strict consensus MP tree differs from the ML tree in the positioning of *Trochodendraceae*. The MP tree placed *Trochodendraceae RPB2* within the *RPB2-d* clade, while the ML tree placed *Trochodendraceae RPB2* and *Buxaceae RPB2* as sister groups before the *RPB2* gene duplication. Neither position is strongly supported. While the tree inferred from the short dataset (exon 11 to 17) is, as would be expected, less strongly supported than that in Fig 2.1(exon 2 to exon 24), the grouping of taxa within the *RPB2-i* and *RPB2-d* clades is the same for both. In the *RPB2* gene tree, Ranunculales splits into two clades: One is composed of *Papaveraceae* (+ *Fumariaceae* and + *Pteridophyllaceae*) and the other the remainder of members of Ranunculales (*Eupteleaceae*, *Berberidaceae*, *Lardizabalaceae*, and *Ranunculaceae*).

Duplication and paralog sorting of the *RPB2* gene occurred in several families within Ranunculales (Fig 2.3). Following an early gene duplication in *Papaveraceae*, most species kept both paralogs, as in the *Eschscholzioidae*, *Glaucium*, *Sanguinaria* and *Chelidonium*. In *Argemone*, *Meconopsis*, and *Papaver*, however, only one of these initial types was retained. Since these three genera are grouped together with strong support (PP=0.99, MP=92%), it is reasonable to infer that the other copy was lost in a common ancestor of these three genera. This early *RPB2* duplication happened only in the *Papaveraceae* not in the *Fumariaceae* and *Pteridophyllaceae*. These *RPB2* gene duplications are inferred to have been rather recent ones.

In the ancestor of the second clade, consisting of *Berberidaceae*, *Lardizabalaceae*, and *Ranunculaceae*, a gene duplication occurred. We have found both duplicate genes only in *Helleborus* (*Ranunculaceae*). More taxon sampling will be required to infer the paralog sorting events in these families. Multiple copies of *RPB2* have also been found in

some species in Caryophyllales (Fig 2.3). All of these duplications, either at the family or subfamily level, are shown by the sequences to have occurred independently of the *RPB2 -i* and *-d* duplication.

2.3.4 Differential expression of RPB2-i *and* RPB2-d *genes in Asterids*

To find out whether both *RPB2* gene paralogs are functional, we carried out RT-PCR using paralog-specific primers on RNA samples from different developmental stages and different organ types. In tomato, RPB2-d is expressed at all developmental stages and in every organ studied (Fig. 2.4 A, C). This suggests that *RPB2-d* is the major functional *RPB2* gene in most tissues. The *RPB2-i* gene is also expressed, but only in floral organs. The expression of *RPB2-i* occurs at the early stages of bud development, then becomes most active in the anther (Fig. 2.4 A, C). *RPB2-i* is the sole or major source of *RPB2* transcripts in pollen (Fig. 2.4 A, C, D). No expression of *RPB2-i* has been detected in sepals, petals or vegetative organs. Since the mRNA of *RPB2-i* and *RPB2-d* is found in both flower buds and anthers, it is necessary to know which of the two genes is the major form expressed in these organs. To determine this, consensus primers that cover the same sequences for both copies were designed and used to amplify both copies from the same RNA sample (Material and methods). The resulting PCR product was digested with TaqI or EcoRI to generate fragments that differ in length for the two copies (Fig. 2.4B). The digests were separated by electrophoresis and stained with ethidium bromide (Fig. 2.4C). As in the initial experiments, *RPB2-d* was found to be the only form expressed in vegetative organs. Only a trace amount of *RPB2-d* was expressed in pollen. No expression of the *RPB2-i* gene was detected in the leaf, stem, sepal and petal. The *RPB2-i* gene is the main one expressed in reproductive organs. In the anther, the extent of expression of *RPB2-i* gene is two to three times that of *RPB2-d*. *RPB2-i* is the main expressed *RPB2* gene in pollen (Fig 2.4D). Also we found little expression of *RPB2-i* in green fruit and only trace amounts in the pistil in these experiments. There is a similar expression pattern of *RPB2* genes in *Nicotiana sylvestris* and *Petunia hybridia.*

As in tomato, the *RPB2-d* gene in *Antirrhinum majus* is expressed in all the organs that have been checked, including the vegetative and reproductive organs. The *RPB2-i* gene is only expressed in floral organs, especially in anthers (Fig 2.5A). But in contrast to the situation in tomato, *RPB2-i* has a low level of expression in sepals, petals and pistils. Both *RPB2-d* and *RPB2-i* are expressed in *Antirrhinum* pollen (Fig. 2.5A). The molar ratio of the transcripts from these two genes was determined as described above for tomato. The major expression pattern is the same in *Antirrhinum* as in tomato. *RPB2-i* is the main expressed form in anthers and pollen, while *RPB2-d* transcripts are the major ones found in other organs. In *Antirrhinum* pollen, nearly equal amounts of the *RPB2-d* and *RPB2-i* transcripts were found (Fig. 2.5B).

The expression of *RPB2-i* and *RPB2-d* in *Rhododendron macrophyllum* was checked at two developmental stages. The first is early flower buds, examined in September of the previous year and the second is the following May, when *R. macrophyllum* blooms. In contrast to the other species examined, expression of *RPB2-i* and *RPB2-d* is detected in both vegetative and reproductive tissues (Fig. 2.5, C-E). However, the molar ratio showed that *RPB2-d* is the major expressed paralog in leaf, petal, and bract, while *RPB2-i* is the main one expressed in anthers and pollen (Fig. 2.5E). There is a shift from mainly *RPB2-d* expression to mainly *RPB2-i* expression in the anther from the early stage to the later stage.

2.3.5 Structural modeling of the residues that differ between RPB2-i *and* RPB2-d

The tomato proteins encoded by *RPB2-i* and by *RPB2-d* both have 60% identity and 73% similarity to the *S. cerevisiae* B140 sequence. We modeled the 3-D structure of tomato paralogs using homology modeling on a yeast pol II template (1i50.pdb, (Cramer et al., 2001)) using SWISS-MODEL (Guex and Peitsch, 1997). The modeled structure can be readily superimposed onto the yeast B140 three-dimensional structure. Structural comparison of the proteins encoded by tomato *RPB2-i* and *RPB2-d* reveals that all of the amino acid differences between them reside on the outer surface of RNA Polymerase II

(Fig. 2.6). The variable residues cluster in the regions of "external 1" and "external 2", in the "protrusion" region and in the "lobe" region (Cramer et al., 2001). None of the variable residues lies in the cavity of the polymerase, at or near the DNA/RNA binding sites and active center.

2.3.6 The phylogeny of RPB1 gene in core eudicots

Only one *RPB1* gene has been found in most species of core eudicots. However, two paralogous *RPB1* genes have been found in *Vitis piasezkii* and *Berberidopsis beckleri*. The DNA sequences between the conserved regions F and G, totaling 724 characters, are used for phylogenetic analysis. Both ML and Bayesian inference resulted in the same toplogy (Fig 2.7). The *RPB1* genes of core eudicots are resolved into two distinct groups: one group includes the *RPB1* genes of Rosids, Caryophyllales, Euasterids I and one paralog of *Vitaceae* and *Berberidopsidaceae*; the other group contains *RPB1* genes of *Trochodendraceae*, *Gunneraceae*, *Dilleniaceae*, Saxifragales, Euasterids II, Ericales, *Aextoxicaceae* and the other paralog of *Vitaceae* and *Berberidopsidaceae*. The statistical support for these groups is high for Bayesian inference, but low for ML bootstrap. With weak support, the time of *RPB1* duplication is found to be subsequent to the divergence of *Buxaceae*.

2.4 Discussion

2.4.1 The RPB2 *Gene Duplications*

In this chapter we have shown that RPB2 gene duplication and paralog sorting events occur in core eudicots. Based on previous work (Oxelman et al., 2004) and this study, we are able to construct the RPB2 gene tree in angiosperms (Fig 2.8). Our studies show that duplication of the RPB2 gene that resulted in RPB2-d and RPB2-i paralogous copies happened at or near the time of the Buxaceae/ Trochodendraceae divergence. Subsequently, the two paralogs evolved as RPB2-i and RPB2-d genes and paralog sorting

of these two genes occurred in different core eudicot taxa. The RPB2-i clade contains the RPB2 genes of Dilleniaceae, Gunneraceae, Vitaceae, Aextoxicaeae , Berberidopsidaceae, and RPB2-i genes of Ericales, Aquifoliales, Escalloniaceae , and Gentianales, Solanales and Lamiales in Asteridae I. The RPB2-d clade contains the RPB2 genes of Saxifragales, Santalales, Caryophyllales, Rosids, Cornales, Asteridae II and the RPB2-d genes of Ericales and Asteridae I. Based on APGII gene tree, we can infer the paralog sorting events in different lineages (Fig 2.9). There are at least 12 independent paralog sorting events after RPB2 gene duplication.

2.4.2 Differential expression of RPB2-i *and* RPB2-d *in the asterids*

We report here differential expression of the second-largest subunit of pol II (*RPB2*) genes in asterids: one copy (*RPB2-d*) is the main *RPB2* gene expressed in vegetative organs, and the other (*RPB2-i*) is in reproductive organs, pollen in particular. Recent studies of pollen transcriptomes shows that there is a striking difference in gene expression profile between pollen and all other tissues and organs in Arabidopsis (Becker et al., 2003; Honys and Twell, 2003). The tissue-specific pattern of expression of duplicate genes for a subunit of the pol II core enzyme raises the following question: Are the pol II core enzymes containing these two proteins isozymes which, by differential association with transcription factors, transcribe genes in a tissue specific manner?

The nonsynonymous/synonymous substitution (dN/dS) ratios, ranging from 0.006 to 0.050 in different lineages, suggest that strong purifying selection acts on both the *RPB2-i* and *RPB2-d* genes (data not shown). All the sites and residues important for RNA polymerase function are conserved between *RPB2-i* and *RPB2-d.* Strikingly, all the variable amino acid residues are on the outer surface of the polymerase. Because paralog sorting happened repeatly in different lineages, we argue that the changes differentiating between the *RPB2-i* and *-d* proteins do not affect the essential catalytic functions of RNA polymerase II. The respective expression patterns of these genes, on the other hand, may provide an explanation for the retention of both gene copies in asterids. Differential

complementary patterns of gene expression between *RPB2-i* and *-d* genes in *Rhododendron* and higher asterids may explain why both genes must be preserved.

2.4.3 Genome duplication and subsequent gene loss as a factor in core eudicot evolution

A comparison of Fig. 2.1 with Fig. 2.7 shows that duplication of RPB1 and RPB2 occurred at approximately the same stage in angiosperm evolution, after the divergence of Buxaceae and close to the time of divergence of Trochodendraceae. While the similarity in these duplication times may be merely a coincidence, with the duplications being independent events, additional considerations suggest otherwise. Within this same time span, other genes, most notably several MADS box genes, also were duplicated. Because MADS box genes are involved in determining floral architecture, their duplication may have been instrumental in the diversification of core eudicots. After divergence of Buxaceae, the group B gene *AP3* duplicated to form the *euAP3* and *TM6* lineages in core eudicots; subsequently, the *TM6* gene was lost in a number of rosids (Kramer et al., 1998). The group C gene *AGAMOUS* and group A gene *APETALA1* each had a duplication prior to core eudicot divergence (Litt and Irish, 2003; Kramer et al., 2004). Genome wide studies of duplicate gene pairs in *Arabidopsis thaliana* indicated a genome duplication predating core eudicot evolution (Bowers et al., 2003). The duplication of so many different gene families at or near the same time suggests that all these events are consequences of a genome duplication event, *i.e.* polyploidization, before core eudicot divergence, at or near the time of the *Buxaceae/ Trochodendraceae* divergence.

Polyploidy is known to occur in a large number of plant families, up to as many as 50% to 70% of them in angiosperms(Masterson, 1994; Wendel, 2000). As compared to their diploid relatives, many tetraploids have better colonizing ability, a higher selfing rate, increased heterozygosity, and increased genetic diversity. These features facilitate the adaptation of polyploids to new ecological niches and may contribute to their overall

reproductive success. Consequently, polyploidy has been considered a major mechanism of adaptation and speciation in plants (Stebbins, 1950, 1971; Grant, 1981; Ramsey and Schemske, 1998; Soltis and Soltis, 2000). However, the evolutionary significance of polyploidy that is, whether it plays an important role in the evolution and diversification of larger lineages, remains an open question (Otto and Whitton, 2000; Crawford and Smocovitis, 2004). This is because of the difficulty in obtaining evidence that polyploidy has changed the rate and pattern of species diversification and because there is no simple way to infer ancient polyploidy by counting either homologous genes or homologous chromosomes. Subsequent evolutionary losses have, in general, wiped out the direct evidence for ancient polyploidy. Our phylogenetic inferences provide evidence of another sort for the existence of an ancient polyploidization event. It is marked by the appearance of two gene lineages where previously there was only one. This happened for a number of genetic loci shortly before the emergence of the core eudicot lineage. Subsequent to the polyploidization, there arose many lineages leading to new plant families and to rapid rise and diversification of core eudicots. Then, in the different diversifying lineages of core eudicots, the *RPB1* and *RPB2* paralogs were sorted independently in the various descendent branches. Thus, our studies support the idea that polyploidization can potentiate evolutionary diversity (Grant, 1981; Otto and Whitton, 2000). We further propose that the stable state of the polyploids and the partial diploidization processes thereafter may have been the driving force behind the remarkable radiation of plant families which occurred early in core eudicot phylogeny. Further investigations of nuclear gene lineages generated in this period should make it possible to assess the impact and association of such a polyploidization event with core eudicot diversification.

2.4.4 Polyploidization and Darwin's abominable mystery of angiosperm evolution

Angiosperms represent one of the greatest terrestrial species radiations of the Neogene. Over 250,000 extant species have been identified. Angiosperms appeared suddenly in the fossil record approximately 130 million years ago (mya) and major

lineages of angiosperms diverged over the next 20 to 40 million years. Charles Darwin described this rapid rise and early diversification of angiosperms as "an abominable mystery" (Darwin, 1903). Since then, many attempts have been made to understand angiosperm diversification, but it still remains one of the most persistent puzzles in modern evolutionary biology (Friis et al., 2005).

Recent studies revealed that the *Arabidopsis* genome underwent three rounds of genome duplication (α, β and γ events, respectively). The β event predated core eudicot evolution and the γ event predated monocot and dicot divergence (Bowers et al., 2003). However, the precise time for the occurrence of these duplication events is unknown. Here we associate the β duplication event with the ancestral polyploidization that, directly or indirectly, gave rise to paralogous gene copies of RPB1, RPB2 and MADS box genes. The phylogenies of these genes place the time of the β duplication at or near the time of the *Buxaceae/Trochodendraceae* divergence, immediately before core eudicot diversification occurred. The fact that polyploidization occurred prior to the species diversification suggests that it provided the genetic potential for species diversification. Indeed, computer simulation of large-scale and small-scale duplication events in *Arabidopsis* suggests that the three genome duplication events are responsible for over 90% of the increase in transcription factors, signal transducers and developmental genes (Maere et al., 2005). We argue here that the β polyploidization event might be directly responsible for the core eudicot radiation and, more generally, that the ancient β and γ genome duplication events may have directly contributed to the rapid rise and early diversification of angiosperms.

Not only did polyploidization provide extra genetic material for evolving new functions and establishing new regulation networks; more importantly the cytogenetic changes it potentiated, such as chromosomal rearrangement, activation of transposable elements, gene inactivation and losses, and epigentic regulation (Ma and Gustafson, 2005), also played an important role in plant speciation. Large chromosomal rearrangements such as reciprocal translocations lead to multivalents formation through

association of nonhomologous chromosomes during meiosis, producing aneuploid gametes. Thus chromosomal rearrangements can act as a post-zygotic mechanism for speciation (Stebbins, 1971; White, 1978). Gene loss has often been considered as a passive fate of most duplicated genes in the polyploids. Here we argue that large scale gene loss and paralog sorting events, resulted either from chromosomal loss/ rearrangements or accumulated mutations overtime, can create partial aneuploids with the result that karyotypic divergence leads to speciation. At the same time, the prolonged processes of paralog sorting events observed for the RPB1 and RPB2 genes suggest that their polyploidy was relatively stable in the ancestors of core eudicots. This prolonged process of polyploid evolution may have been important for generating the genetic diversity and variation over evolutionary time that led to angiosperm diversification. We thus further propose that the stable state of the polyploids and the slow evolutionary processes of polyploids thereafter may have been the driving force behind the remarkable radiation of plant families which occurred early in core eudicot phylogeny.

Table 2.1. List of taxa and voucher specimens studied in this dissertation.

Higher taxon	Family	Species	Genbank #	Voucher/Reference
	Nymphaeaceae	*Nymphaea colorata*	AF043427, DQ058634	Yoshikawa349347
MAGNOLIIDS				
Laurales	Calycanthaceae	*Chimonanthus praecox*	DQ017093	
	Lauraceae	*Lindera glauca*	DQ017116	Yoshikawa 349352
Magnoliales	Magnoliaceae	*Liriodendron tulipifera*	DQ058631	Yoshikawa 349354
MONOCOTS				
Dioscoreales	Dioscoreaceae	*Dioscorea sansibarensis*	AY563268	
EUDICOTS				
	Buxaceae	*Buxus sempervirens*	DQ017091; RPB1:DQ228258	Yoshikawa 349374
		Sarcococca hookeriana	DQ017130	
	Trochodendraceae	*Trochodendron araloides*	AY563269	Yoshikawa 349360
	Tetracentraceae	*Tetracentron sinense*	DQ017132; RPB1:DQ228281	Yoshikawa 349353
Proteales	Nelumbonaceae	*Nelumbo nucifera*	DQ017120; RPB1:DQ228273	
	Proteaceae	*Embothrium coccineum*	DQ017103, DQ017104	Yoshikawa 349339
	Platanaceae	*Platanus orientalis*	DQ058632; RPB1a: DQ228272;RPB1b:DQ 228271	Yoshikawa 349338
		Macadamia integrifolia	RPB1a:DQ228269; RPB1b:DQ228270	
Ranunculales	Berberidaceae	*Mahonia nervosa*	DQ017117	Yoshikawa 349355
		Nandina domestica	DQ017119	Yoshikawa 349366
	Eupteleaceae	*Euptelea polyandra*	DQ017105	Yoshikawa 349382

Table 2.1 (continued)

	Lardizabalaceae	*Akebia longiracemosa*	AY566614	Yoshikawa 349380
	Papaveraceae	*Argemone mexicana*	DQ017088 (1); DQ017089 (2)	Yoshikawa 349356
		Meconopsis betonicifolia	DQ017118	Yoshikawa 349343
		Papaver somniferum	DQ017122	
		Chelidonium majus	DQ017094 (1); DQ017095 (2)	Yoshikawa 349367
		Glaucium flavum	DQ017106 (1); DQ017107 (2)	Yoshikawa 349341
		Sanguinaria canadensis	DQ017128 (1); DQ017129 (2)	Yoshikawa 349373
		Eschscholzia californica	AY566616 (1); DQ017110 (2)	Yoshikawa 349368
	Fumariaceae	*Corydalis flexuosa*	DQ017086 (1); DQ017098 (2)	Yoshikawa 349381
		Dicentra formosa	DQ017100	Yoshikawa 349362
	Pteridophyllaceae	*Pteridophyllum racemosum*	DQ017125	Oxelman 2235
	Ranunculaceae	*Helleborus orientalis*	DQ017109 (1); DQ017110(2)	Yoshikawa 349342
		Hydrastis canadensis	DQ017115	Oxelman 2249
CORE EUDICOTS				
	Aextoxicaceae	*Aextoxicon punctatum*	DQ017087; RPB1: DQ228254	73.0753 (UCBG)
	Berberidopsidaceae	*Berberidopsis beckleri*	DQ020634; RPB1a :DQ228255; RPB1b: DQ228256	Yoshikawa 349384*
		Berberidopsis corallina	DQ020633; RPB1: DQ228257	Yoshikawa 349383*
	Dilleniaceae	*Dillenia suffruticosa*	DQ017101; RPB1: DQ228260	95.1119 (UCBG)
		Hibbertia aspera	DQ017111; RPB1:DQ228264	Yoshikawa 349371*
		Hibbertia scandens	DQ017113; RPB1:DQ225265	87.1455 (UCBG)
		Hibbertia cuneiformis	DQ017112; RPB1: DQ228263	Yoshikawa 349372*
Gunnerales	Gunneraceae	*Gunnera chilensis.*	DQ017108; DQ094142; RPB1:DQ228262	
Caryophyllales	Amaranthaceae	*Spinacia oleracea*	AF020840, DQ058635	

Table 2.1 (continued)

		Iresine herbstii	RPB1: DQ228267	
	Aizoaceae	*Ruschia herrei*	RPB1: DQ228277	
	Cactaceae	*Pereskia aculeata*	DQ017123 (1); DQ017124 (2)	Yoshikawa 349365
	Caryophyllaceae	*Dianthus caryophyllus*	DQ017099	Yoshikawa 349340
	Droseraceae	*Drosera capensis*	DQ017102	Yoshikawa 349364
		Dionaea muscipula	RPB1: DQ228261	
	Nepenthaceae	*Nepenthes maxima superba*	DQ017121	Yoshikawa 349369
		Nepenthes ampullaria	RPB1: DQ228274	
	Phytolaccaceae	*Rivina humilis*	DQ017126 (1); DQ017137 (2)	Yoshikawa 349361
	Plumbaginaceae	*Armeria maritima*	DQ017085 (1); DQ017090 (2)	Yoshikawa 349379
	Polygonaceae	*Homalocladium platycladum*	DQ017114	Yoshikawa 349370
Santalales	Loranthaceae	*Arceuthobium campylopodium*	AY566624	Yoshikawa 351264
Saxifragales	Hamamelidaceae	*Liquidamber acalycina*	AY566623, DQ058633	Yoshikawa 350770
	Saxifragaceae	*Astilbe* sp.	DQ058629	Yoshikawa 349385
	Crassulaceae	*Kalnchoe tomentosa*	RPB1: DQ228268	
		Sedum burrito	RPB1a: DQ228278; RPB1b:DQ228279	
ROSIDS				
	Vitaceae	*Cissus tuberosa*	DQ017096; RPB1: DQ228259	
		Vitis piasezkii Maxim.	DQ017133; RPB1a :DQ228282; RPB1b:DQ228283	Yoshikawa 349358
Myrtales	Myrtaceae	*Callistemon subulatus*	DQ017092	Yoshikawa 349337
EUROSIDS I				
Cucurbitales	Coriariaceae	*Coriaria sarmentosa*	DQ017097	Yoshikawa 349375
	Cucurbitaceae	*Kedrostis africanca*	RPB1: DQ228276	

Table 2.1 (continued)

		Ibervillea sonorae	RPB1: DQ228266	
Fagales	Fagaceae	*Quercus* coccinea	RPB1: DQ228275	
EUROSIDS II				
Brassicales	Brassicaceae	*Arabidopsis thaliana*	Z19121	Larkin and Guilfoyle, 1993
Sapindales	Sapindaceae	*Acer palmatum var. amoenum*	DQ071432; RPB1: DQ228253	Yoshikawa 349378
ASTERIDS				
Ericales	Diapensiaceae	*Diapensia lapponica*	AY579381 (D); AY579382 (I)	Gage 1225
	Ericaceae	*Rhododendron macrophyllum*	DQ058627 (I); DQ058628 (D) ; RPB1: DQ020635	Yoshikawa 351265
	Polemoniaceae	*Linanthus californicus*	DQ058636 (I); DQ058637 (D)	Yoshikawa 349351
	Sarraceniaceae	*Sarracenia alata*	DQ017131	Yoshikawa 349345
	Theaceae	*Camellia japonica*	AY566627, DQ058625 (D); AY566628, DQ058626 (I)	NY25
EUASTERIDS I				
Gentianales	Rubiaceae	*Gardenia* sp.	AJ558243 (I); AJ558358 (D), AJ558359 (D)	Oxelman 2319
Lamiales	Scrophulariaceae	*Antirrhinum majus*	DQ020637 (I); DQ020642 (D) RPB1: DQ020643	
Solanales	Solanaceae	*Petunia hybrida*	DQ020638 (I); DQ020641 (D)	
		Solanum lycopersicum	DQ020639 (I); RPB1: DQ020644	
		Nicotiana sylvestris	DQ020636 (I); DQ020640 (D)	
EUASTERIDS II				
Asterales	Asteraceae	*Senecio picticulis*	RPB1: DQ228280	

*: Strybing Arboretum & Botancal Gardens UCBG: University of California, Botanical Garden

Table 2.2 Results of Shimodaira-Hasegawa (SH) and parametric bootstrap tests for alternative hypotheses.

	SH test				Parameric bootstrap test			
Hypothesis	-lnL	δ -lnL	SH	rejected*	Length	δlength	P	rejected*
ML tree	37378.375	----	----	Best	----	-----	----	----
MP tree	-----	----	----	-----	8329	-----	----	Best
Tetracenton diverged before duplication	37388.179	9.805	0.567	No	8335	6	0.14	No
Tetracenton RPB2 in the RPB2-I clade	37388.378	10.003	0.555	No	8335	6	0.09	No
Dilleniaceae RPB2 in RPB2-D clade	37408.714	30.339	0.077	No	8348	19	0.01	Yes
Gunnera RPB2 in RPB2-D clade	37420.426	42.051	0.011	Yes	8377	48	<0.01	Yes
Vitaceae RPB2 in RPB2-D clade	37464.656	86.281	<0.001	Yes	8375	46	<0.01	Yes
Proteales RPB2 in RPB2-D clade	37385.686	7.311	0.649	No	8334	5	0.18	No
Proteales RPB2 in RPB2-I clade	37384.918	6.543	0.645	No	8334	5	0.03	Yes

* hypothesis is rejected as significantly different for both tests if P-Value < 0.05.

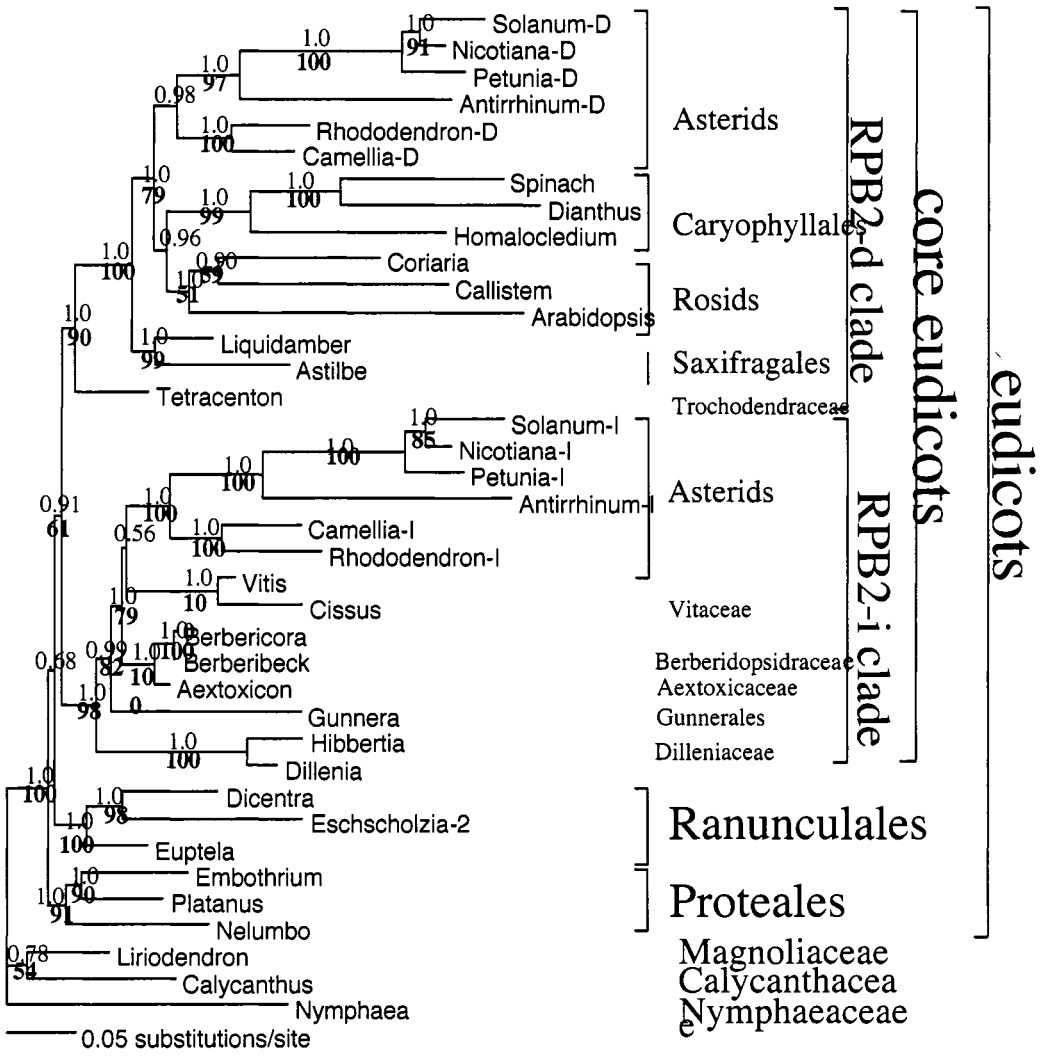


Figure 2.1 The ML RPB2 gene tree. The summary tree from Bayesian inference resulted in the same tree topology. Posterior probability of Bayesian inference is shown above branches. Bootstrap of ML over 50% is shown below the branches.

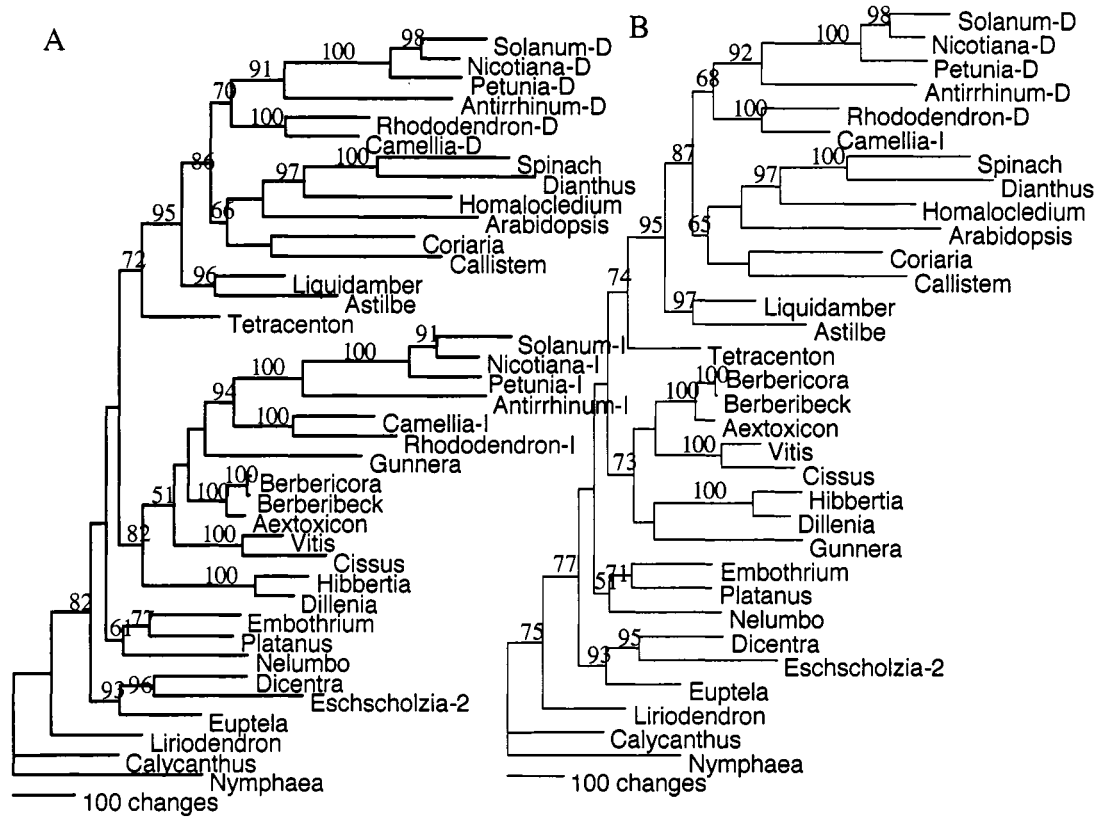


Figure 2.2 A. The MP RPB2 gene tree. B. The MP RPB2 gene tree after deletion of all the RPB2-I genes of Asterids. MP bootstrap over 50% is shown above the branches.

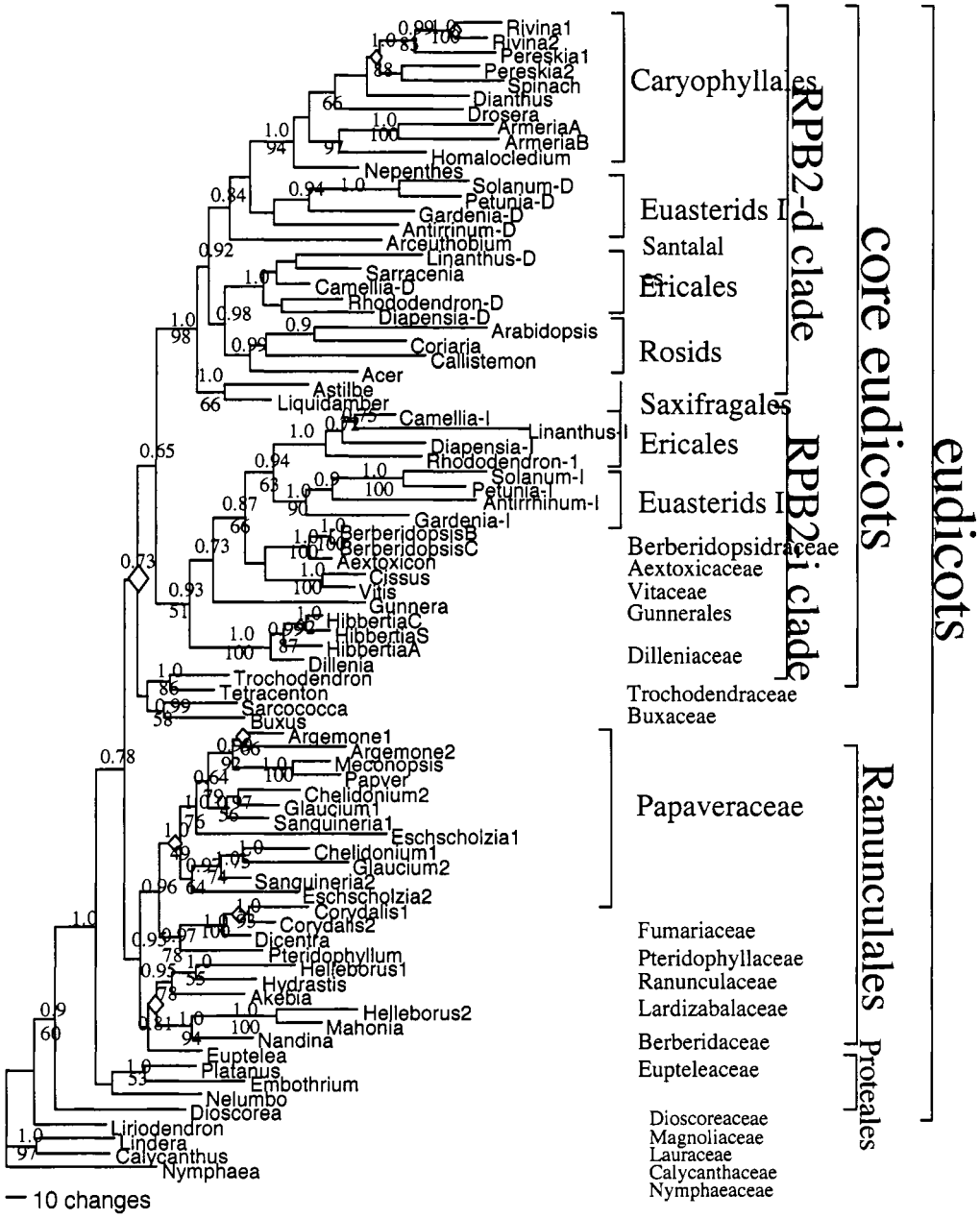


Figure 2.3 The RPB2 ML gene tree using exon 11 to exon 17. Bayesian inference has the same tree topology as ML. The posterior probability of Bayesian inference is shown on the top of the branches and the bootstrap of MP over 50% is shown underneath the branches. Possible duplication events are marked as ◆.

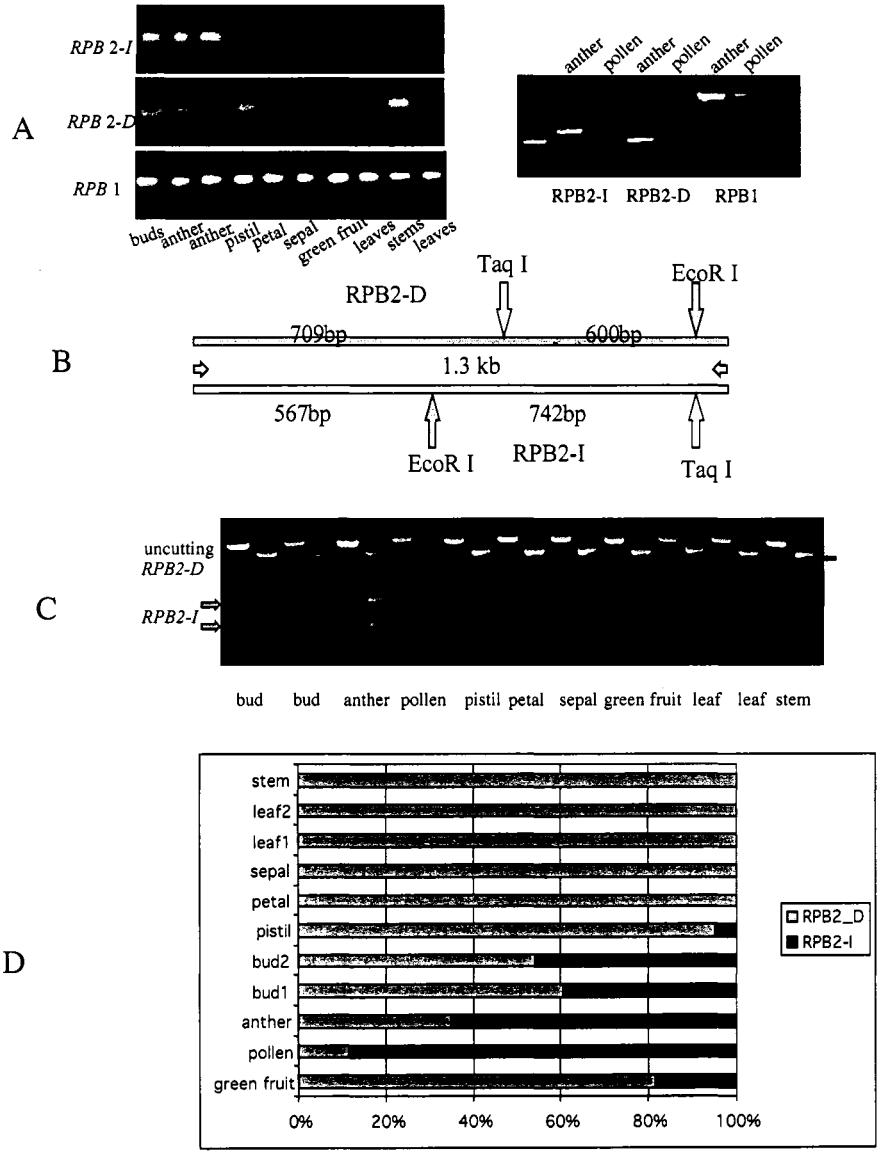


Figure 2.4 Expression of RPB2-i andRPB2-d in *Solanum lycopersicon.* ***A.*** Expression of RPB2-i and RPB2-d in different tissues by RT-PCR. B. Schematic illustration of enzymatic digestion of RT-PCR product to distinguish the two RPB2 cDNAs in *Solanum lycopersicon.* C. A picture of digestion pattern by Taq I. D. The molar ratio of RPB2 genes in tomato. The ratio of D/I in buds is about 2.56 ± 0.54; the ratio of I/D in anther is about 3.25 ± 0.32.

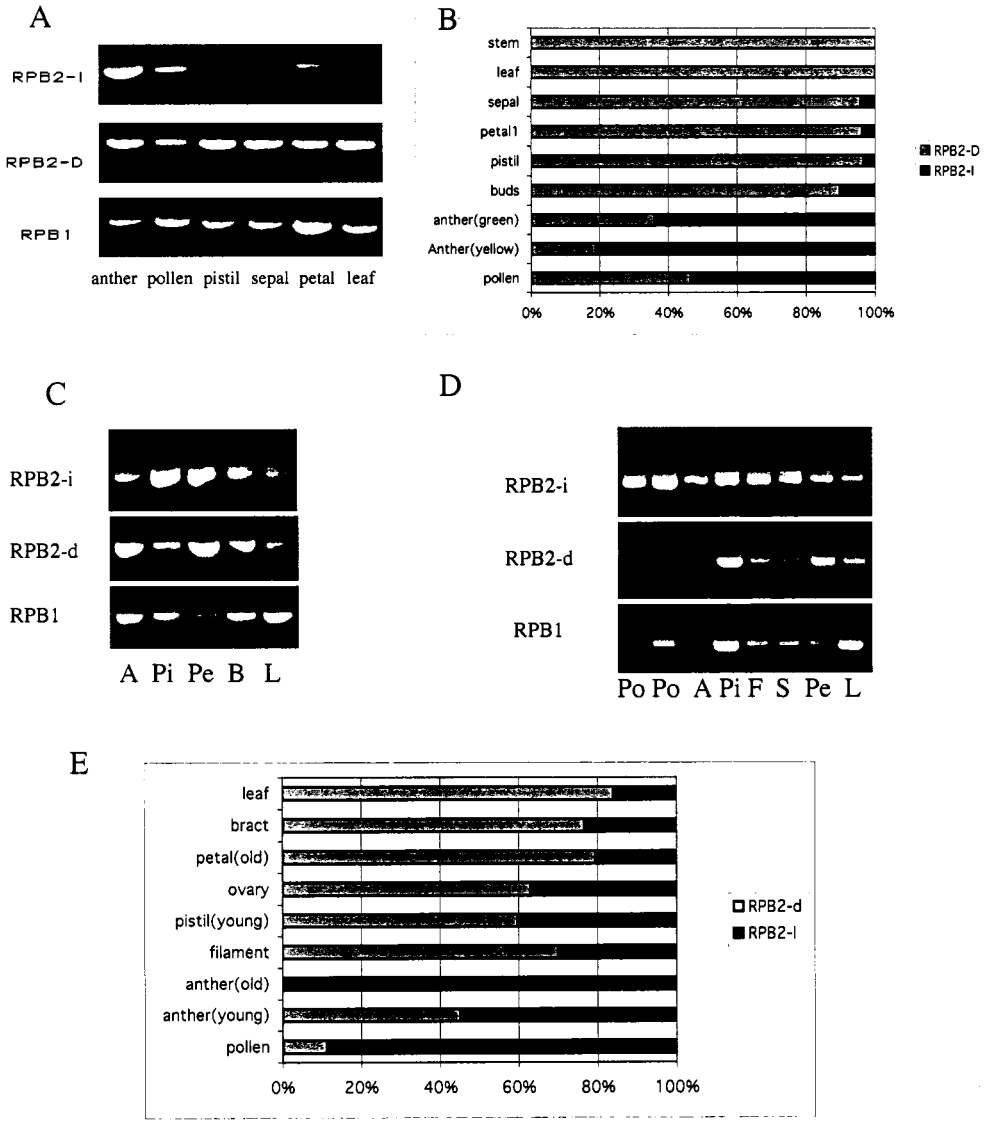


Figure 2.5 A. Expression of RPB2-d and RPB2-i in *Antirrhinum majus*. B. The molar ratio of RPB2-d and RPB2-i in different organs in *Antirrhinum majus* as determined by Ban I. C-E. Expression of RPB2-d and RPB2-i in *Rhododendron macrophyllum.C:* In budding stage. D. In flowering stage. E. The molar ratio of RPB2-i and RPB2-d in *Rhododendron macrophyllum* cutted by BbsI. A: anther; Pi: pistil; Pt: petal; F: filament; S: style; B: bract; L: leave. Young refers to the early stage. Old refers to the flowering stage.

Figure 2.6 Homology modeling of RPB2-I and RPB2-D proteins in tomato. The RPB2-I subunit is shown in the picture as green and red ribbon. The other subunits of yeast pol II are shown as white ribbons (left). The red ribbons are the residues that haven't been resolved in the yeast pol II structure (1i50.pdb). The green ribbons are the corresponding residues which can be superimposed onto the yeast RPB2 structure. The residues that differ between tomato RPB2-I and RPB2-D are shown as filled balls on the left figure. On the right, various protein domains are highlighted in different colors on the modeled RPB2-i protein. dark blue: external region1; red: external region 2; light blue: Lobe region; yellow; protrusion.

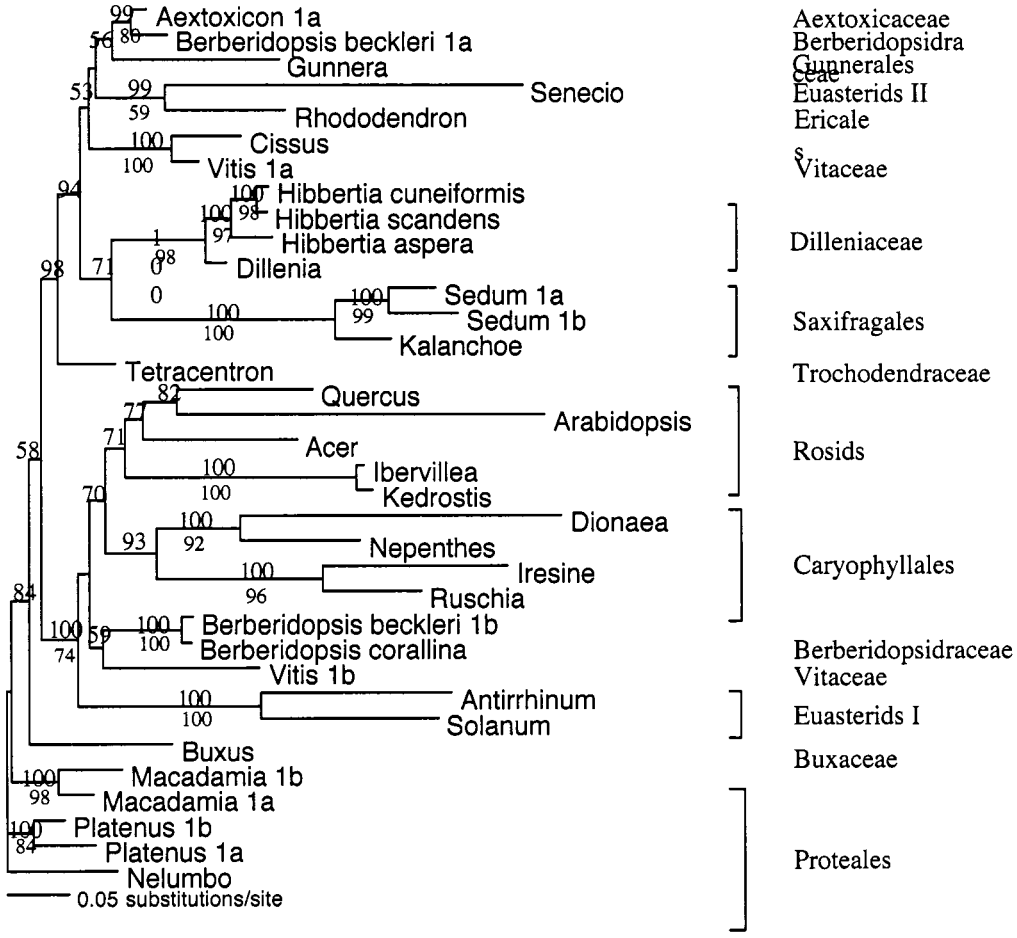


Figure 2.7 The ML RPB1 gene tree. Posterior probability of Bayesian inference is shown above branches. Bootstrap of ML over 50% is shown below the branches.

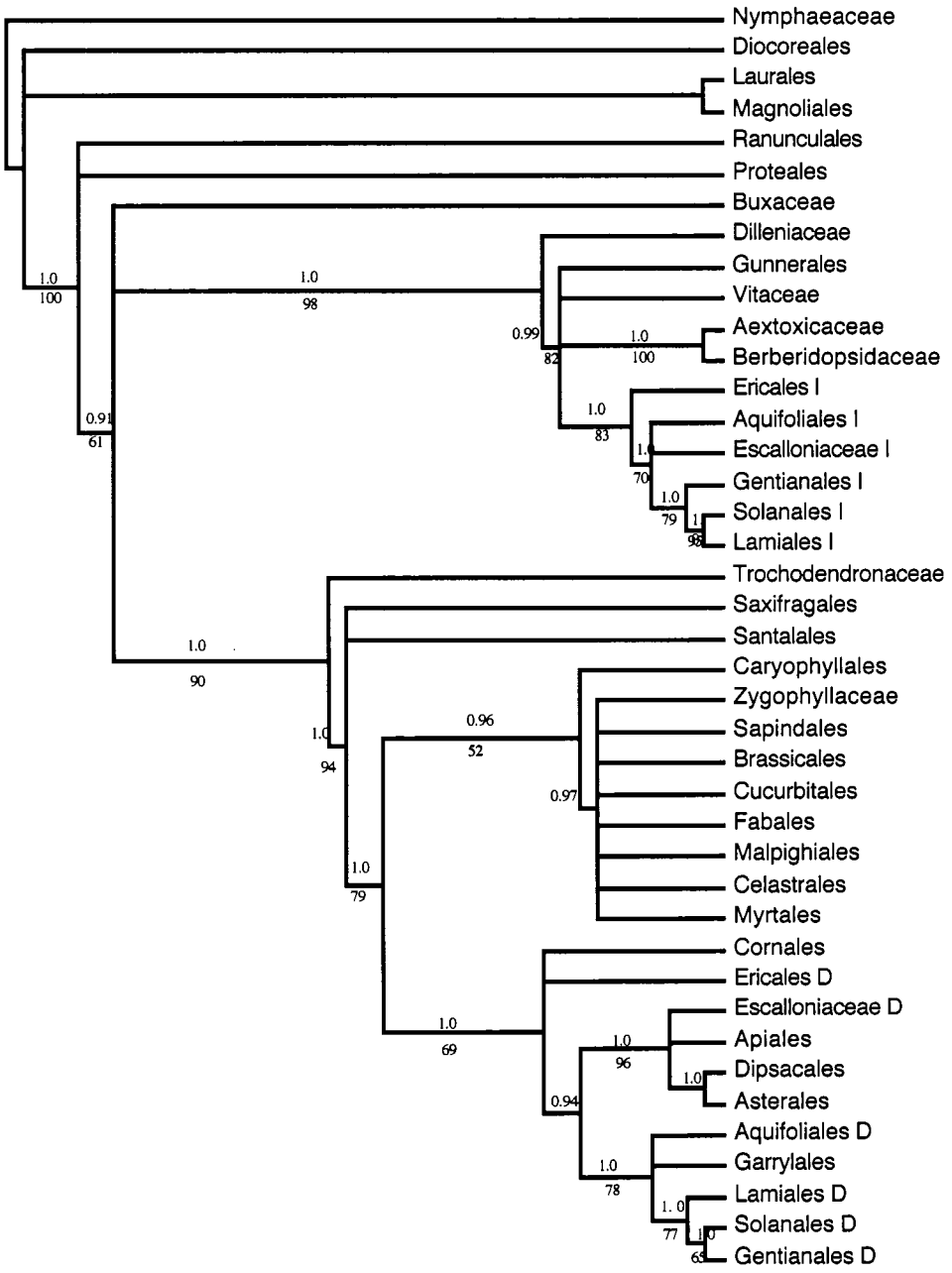


Figure 2.8 The summary RPB2 gene tree in angiosperms. The bayesian inference support is shown above the branches. ML bootstrap support is shown under the branches.

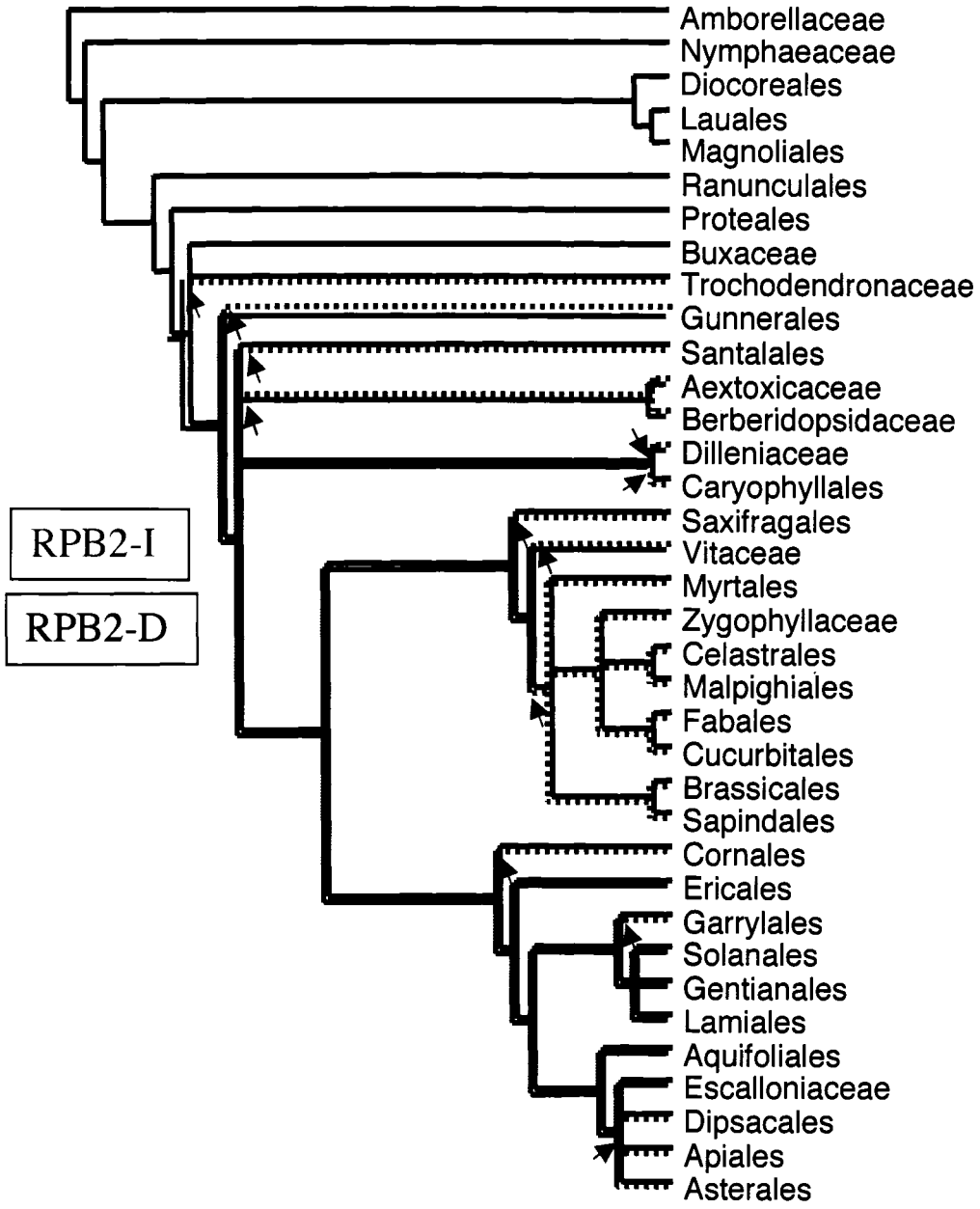


Figure 2.9 The RPB2 gene duplication and gene loss events in core eudicots. The APGII is used as species tree. The dashed lines represent the gene loss events. The arrows indicate the gene loss events in different lineages.

Chapter 3. A multi-step process gave rise to RNA polymerase IV of land plants

Summary

Since their discovery in Metazoa, the three nuclear RNA polymerases (RNAPs) have been found in fungi, plants and diverse protists. In all eukaryotes studied to date, RNAPs I, II and III collectively transcribe all major RNAs made in the nucleus. We have found genes for the largest subunit (RPD1/RPE1) of a new DNA-dependent RNA polymerase, RNAP IV, in all major land plant taxa and in closely related green algae and genes for the second-largest subunit (RPD2) of this enzyme in all land plants. Phylogenetic study indicates that RNAP IV genes are sister to the corresponding RNAP II genes. Our results show the genesis of RNAP IV to be a multistep process in which the largest and second-largest subunit genes evolved by independent duplication events in the ancestors of Charales and land plants. These findings provide insights into evolutionary mechanisms that can explain the origins of multiple RNA polymerases in the eukaryotic nucleus.

Key words: RNA polymerase IV, RPD1, RPD2

3.1 Introduction

DNA dependent RNA polymerase (RNAP) is the principal enzyme responsible for gene transcription, the first step of gene expression and the major target of developmental and environmental regulation. There are two categories of RNAPs: single subunit enzymes, exemplified by T7 RNAP and multiple subunit cellular RNAPs. In plants, nuclear encoded single subunit RNAPs are responsible for the transcription of mitochondrial genes and some plastid genes. (Hedtke et al., 1997; Hess and Borner,

1999). Nuclear and chloroplast multiple subunit RNAPs, containing 5 to 15 subunits, are responsible for transcription of their respective genomes. Green plants, like metazoans, fungi, and all protists studied, have three nuclear DNA dependent RNA polymerases (RNAPs). RNAP I transcribes rRNA; RNAP II transcribes mRNA and most of sn RNA; and RNAP III transcribes tRNA, 5s rRNA and some sn RNAs. Each nuclear RNA polymerase has a unique pair of largest and second largest subunits, analogous to the eubacterial β' and β proteins, as well as 8 to 10 smaller subunits, some of which are shared (Sentenac, 1985; Young, 1991). The two largest subunit genes are named, respectively, RPA1 and RPA2 for RNAP I, RPB1 and RPB2 for RNAP II and RPC1 and RPC2 for RNAP III. Archaebacteria have a single RNAP with extensive subunit similarity to the eukaryote nuclear RNAPs (Langer et al., 1995; Bell and Jackson, 1998).

The largest and second largest subunits of all multiple subunit polymerases contain the amino acid residues for the active center, for template DNA binding and for DNA/RNA hybrid binding. These sequences are conserved throughout evolution (Mooney and Landick, 1999; Korzheva et al., 2000; Cramer et al., 2001). There are 8 conserved regions in all the largest subunits, named regions A through H (Jokerst et al., 1989; Puhler et al., 1989). There are 9 conserved regions in all the second largest subunits, named regions A through I (Sweetser et al., 1987). The three dimensional structures have been determined for RNAP from *Thermus aquaticus* (Zhang et al., 1999), *Thermus thermophilus* (Vassylyev et al., 2002) and for *S. cerereisae* RNAP II (Cramer et al., 2001; Gnatt et al., 2001). The conserved regions have proven either to make up the enzyme active center or to be important for subunit assembly (Cramer et al., 2001). RPB1 and RPB2 gene sequences have been widely used in phylogenetic study of lower eukaryotes, fungi and plants (Stiller and Hall, 1997; Stiller et al., 1998; Liu et al., 1999; Dacks et al., 2002; Liu and Hall, 2004).

Three DNA dependent RNA polymerases have been purified from animals (Roeder and Rutter, 1969; Greenleaf and Bautz, 1975; Sklar et al., 1976), yeasts (Adman et al., 1972; Young and Whiteley, 1975), plants (Jendrisak and Burgess, 1975; Guilfoyle

et al., 1976) and protists (Pong and Loomis, 1973). Genome projects have shown that there are three and only three sets of genes coding for RNA polymerase largest and second largest subunits in the nuclear genomes of fungi, metazoa, and various protists as well as the nucleomorph of the cryptomonad *Guillardia theta* (Douglas et al., 2001). In plants, the Arabidopsis Genome Initiative revealed four genes in the families encoding both the largest and second largest subunits of nuclear DNA directed RNA polymerases (AGI, 2000). This has proven to be also true in the rice genome (Goff et al., 2002; Yu et al., 2002). Distinct from the genes for RNAP I, RNAP II and RNAP III, there is another set of genes coding for a new putative RNA polymerase. Here we report the phylogenetic study of these genes.

3.2 Methods and materials

3.2.1 PCR and cloning

In order to obtain RNAP IV coding sequences, we first used the four all-polymerase primers to amplify regions of RPD1, RPE1 and RPD2 (Fig. 3.1). Subsequently, the rest of each coding sequence was recovered by RT-PCR and/or RACE. For the largest subunits, the primers were RP-Dfor: CCCTACAATGCIGAYTTYGAYGGIGA; RP-Erev: GGTCCGAAIAIYTGYTTICCIGTCCA. The PCR condition was set as: 5 cycles of 94 °C, 1min; 45 °C 1min; increase to 72°C at a slope of 1 °C per 5 sec.; 72 °C 1min. followed by 30 cycles of 94 °C 1min, 55 °C 1min, 72 °C 1min. For the second-largest subunit, the primers were RNAP-10F: TTYTCIAGYATGCAYGGICARAARGG; RNAP11R: ARRCARTCNCKYTCCATYTCNCC. The PCR condition was set for 35 cycles of 94 °C, 1min; 50°C 1min; increase to 72°C at a slope of 1 °C per 5 sec.; 72 °C 1min. The PCR products were directly cloned into TOPO-TA vectors (Invitrogen) for sequencing. The taxon information and gene bank accession numbers for the genes studied are provided in Table 3.1.

3.2.2 Phylogenetic Analyses

The protein sequences were first aligned in ClustalX. Then the alignment was adjusted using Se-Al 2.0 (Rambaut, 1996). Maximum likelihood (ML) analysis was carried out by PHYLIP 3.6 (Felsenstein) under JTT model and six categories of gamma distribution for among-site variations. Bootstrap for ML was done for 100 replicates. Bayesian inference was conducted by MrBayes 3.0 (Ronquist and Huelsenbeck, 2003) under the JTT model and six categories of gamma distribution for among-site variation. We ran four chains for 10^6 generations and sampled every 100 generations and the analysis was done three times independently. The trees were summarized by 50% majority rule consensus after an initial 2,000 trees "burn-in".

The relative rate of evolution between orthologs of RNAPs in *Arabidopsis* and *Oryza* is estimated by the number of substitutions per site, d, by numerically solving the equation $q= \ln(1+2d)/2d$, where q is the proportion of identical sites between aligned sequences. This equation accounts for the general case where substitution rates can differ for both amino acids and sites (Grishin, 1995).

3.3 Results

3.3.1 The genes for the largest and second-largest subunits of RNA polymerases in the nuclear genomes of eukaryotes

Using the protein sequences of the largest and second largest subunits of RNAP II to perform blast searches against the complete sequenced genomes, we can readily identify the genes coding for the largest and the second-largest subunits of RNAPs. There are three and only three largest (RPA1, RPB1, RPC1) and second-largest (RPA2, RPB2, RPC2) nuclear RNAP subunits for fungi, animals and protists (Table 3.2).

2.3.2 The RNA polymerase genes in Arabidopsis and Oryza nuclear genomes

The Arabidopsis Genome Initiative (AGI) first reported four sets of genes coding for the largest and second largest subunits of nuclear RNA polymerases (AGI, 2000). However the annotations of these genes contained several errors. In table 3.3, we have re-annotated the genes coding for the largest and second largest subunits of each RNA polymerase in the nuclear genome. For RNAP I, RNAP II, and RNAP III, there is a single gene for each enzyme's largest subunit and second largest subunit. Each of them has strong homology to its orthologs in other organisms such as yeasts and mammals. There are two genes coding for additional largest subunits (named RPD1 and RPE1) and two genes coding for additional second largest subunits (named RPD2a and RPD2b).

The RPD1(At1g630200) gene was correctly predicted by the AGI annotation. RPE1 was incorrectly annotated as two independent transcripts (At2g40030, At2g 40040), owing to misassignment of one intron position, creating a premature stop codon. Using RT-PCR, we found that both of these sequences are on a single transcript. Both RPD1 and RPE1 are expressed in leaves and flowers. We have sequenced the full length of the RPD1 and RPE1 transcripts and confirmed all intron positions. The intron positions are the same for RPD1 and RPE1, except that intron 8 has been lost in RPE1. Five intron positions in the N-terminal region are shared with Arabidopsis RPB1 (Fig. 1). There is one intron in the 5'UTR for both the RPD1 and RPE1 genes. *A. thaliana* has two genes coding for the second largest subunit of pol IV. RPD2a (At3g18090) and RPD2b (At3g23780) have 96% sequence identity, indicating a recent duplication. Both genes are expressed. However, RPD2a is a psudogene due to a frame-shift mutation in exon 2. There are seven introns in both RPD2 genes, including one intron in the 5' UTR (Fig 3.1). None of these intron positions are the same as those of Arabidopsis RPA2, RPB2 or RPC2.

In the *Oryza sativa* ssp. *japonica* genome, sequenced by The International Rice Genome Sequencing Project (IRGSP), and in the *Oryza sativa* L. ssp. *indica* genome, shotgun sequenced by the Chinese Academy of Sciences (Goff et al., 2002; Yu et al., 2002), the sets of RNA polymerase genes are the same. There is one gene for each of the largest and second largest subunits of RNAP I and RNAP III and one gene for RPB2 (Table 3). There are three genes with very similar sequences coding for RPB1. There are two genes for RPD1, two genes for RPE1, and two genes for RPD2 (Table 3). The two RPD1 genes have 88% sequence identity. The predicted amino acid sequences for the two RPE1 genes have only 42% sequence identity. The intron positions of Oryza RPD1 and RPE1 are identical to those of the Arabidopsis RPD1 and RPE1 genes. The two Oryza RPD2 genes have 80% sequence similarity and the intron positions are the same as those in Arabidopsis.

3.3.3 The substitution rates of RNAP genes in Arabidopsis *and* Oryza

The amino acid sequences of the conserved regions are used to calculate the relative substitution rates (number of substitutions per site) of RNAP orthologs in *Arabidopsis* and *Oryza* (Table 3.3). For each polymerase, the largest and second largest subunit genes have similar substitution rates. RPB1 and RPB2 are the most conserved genes as compared to the largest and the second largest subunit genes of RNAP I, RNAP III and RNAP IV. RPD1/RPE1 have the highest substitution rates, approximately 20 times faster than that of RPB1. The substitution rate of RPD2 is about 10 times faster than that of RPB2.

3.3.4 Amplification of RPD1/RPE1 and RPD2 genes

In order to search for RNAP IV genes in other taxa, we devised a global polymerase gene amplification (PCR) procedure for largest and second-largest subunits. This PCR strategy takes advantage of regions conserved at the protein level in *all* nuclear RNA polymerases, including RNAP IV. For the largest subunit, we used the active site

in conserved region D (NADFDGD) as forward priming site and conserved region E (WTGKQ) as reverse priming site. For the second largest subunit genes, we used the conserved region H (GDKFSSR/MHGQKG) and conserved region I (GEMERD) as priming sites. The primers were designed to include all possible degeneracy. To demonstrate the efficacy of these primers, we successfully amplified all three largest subunit and second largest subunit genes for RNA polymerase I, II and III in *S. cerevisiae* and *Mucor hiemalis*(data not shown).

For the largest subunit, we were able to amplify the RPD1 and RPE1 genes in *Solanum*(asterid), and RPD1 genes in *Cerapteris* (fern), *Anthocerus* (hornwort), *Sphagnum* (moss), *Marchantia* (liverwort), and in *Chara* and *Nitella* (Charales). No RPD1 gene was found in two other close algal relatives of land plants, Coleochaetales and Zygnematales. For the second largest subunit, we were able to amplify RPD2 genes in all land plants tested, including liverworts, mosses, ferns and seed plants. No RPD2 genes were found in green algae.

3.3.5 Phylogeny of RPD1/RPE1 and RPD2 genes

For the RNAP IV largest subunit, all the conserved regions (A through H) are present and can readily be identified. However, the sequences in the conserved regions of RNAP IV are more variable than those of RNAPs I, II and III. In relation to the canonical largest subunit protein structure, all of the RPD1/RPE1 genes have a complete deletion of about 190 amino acids between region F and G, corresponding to the "foot" domain in the yeast RNAP II structure (Cramer et al., 2001) (Fig 3.1).

Phylogenetic analyses were carried out using inferred amino acid sequences of all largest and second largest subunits of eukaryotic RNAPs I-IV. The protein alignments were constructed with sequences from conserved regions A through H of the largest subunits(Jokerst et al., 1989; Puhler et al., 1989) and conserved regions A through I of the second-largest subunits(Sweetser et al., 1987). Because many parts of the largest

subunit sequences are highly divergent between polymerase families, their alignment included only the most conserved domains, totaling 479 characters. Both ML and Bayesian inference resulted in the same tree topology (Fig 3.2A). For each class of RNA polymerase, the largest subunits form a monophyletic group with strong support. The Bayesian posterior probability (PP) for each group is 100%, while ML bootstrap support is 100% for the RNAP I, III, and IV clades, and 87% for RNAP II. The RPD1/RPE1 clade is recovered as sister to the RPB1 clade with moderate support (ML=69%, PP=98%). Using an alignment of 564 positions for second largest subunits, both ML and Bayesian inference result in the same tree topology (Fig 3.3). Second-largest subunits of each of the four polymerases individually form monophyletic groups with strong support. RPD2 genes are recovered as a clade that is strongly supported as sister group to RPB2, (PP=100%, ML=92%).

In angiosperms, there are two variants of the largest subunit genes, RPD1 and RPE1. In view of the clear evidence (Fig. 3.2a) that the RPD1/RPE1 genes of *Arabidopsis* and *Oryza* constitute a monophyletic group exclusive of all other polymerase genes, we sought to compare them with the RPD1 sequences from a broad variety of green plants through phylogenetic analysis of 537 inferred amino acid sequences between regions D and H (Fig 3.2B). The resulting tree generally reflects accepted species relationships(Pryer et al., 2001). Because only one RNAP IV largest subunit gene was found in the Charales, mosses, and ferns, we infer that the gene duplication that resulted in RPD1 and RPE1 occurred after the origin of land plants, but before the diversification of angiosperms.

3.4 Discussion

In this report, we have studied the evolution of genes encoding a putative new RNA polymerase (RNAP IV) in plants. We have found genes for the largest subunit and second largest subunits of RNAP IV in all terrestrial plants and the largest subunit gene in the *Charales*. Phylogenetic analysis indicated that RNAP IV originated from RNAP II.

3.4.1 The genesis of RNAP IV

As is true of all completed eukaryotic genomes outside the green plant lineage, *Chlamydomonas reinhardtii* contains three and only three sets of largest and second-largest subunit genes of nuclear RNA polymerases; these correspond to RNAPs I, II and III (Table 3.2). To localize the origin of RNAP IV during the evolution of green plants, we have tried repeatedly to amplify RPD1 and RPD2 genes in the three closest green algal relatives of land plants: the Charales, Coleochaetales, and Zygnematales (Karol et al., 2001). Authentic RPD1 sequences were obtained from *Chara* spp. and *Nitella* spp. in the Charales, but no verifiable RPD2 gene was found. No RNAP IV genes could be amplified from four species of *Coleochaete*: *C. scutata*, *C. nitellarum* (SAG 3.91), *C. soluta* or *C. irregularis*. Similar negative results were obtained from *Spirogyra* sp., *Zygnema cylindrium* (SAG 689-2), and *Mesotaenium endlicherianum* (SAG 12.97) of the Zygnematales. Because the sequences chosen for PCR priming are generally conserved throughout RNA polymerases I-IV, and because they successfully recovered RNAP IV genes from both land plants and Charalian algae, we interpret these results to mean that RPD1 and RPD2 do not exist in Coleochaetales and Zygnematales. The absence of RPD2 from Charales implies that, for the RPD1 subunit to be in an active enzyme, it would most likely assemble with RPB2 protein in these algae. This assumes that RPD1 arose by gene duplication of RPB1, as suggested by our phylogenic results.

The existence of RNAP IV subunit genes both in Charales and in land plants agrees with recent findings that Charales are the sister taxon to all land plants(McCourt et al., 2004). Based upon this evidence and our results, we suggest that the initial largest subunit duplication giving rise to RNAP IV occurred after the common ancestor of Charales and land plants diverged from other green algae. This hypothesis predicts that phylogenetic analyses should recover the RPD1/RPE1 clade as sister to RPB1 from the Charales and land plants, and the RPD2 clade as sister to land plant RPB2. In other words, RNAP IV genes should nest within the RPB1 green plant phylogeny. For both of

the polymerase subunit trees (Fig 3.2, 3.3), however, RNAP IV genes are sister to the entire RPB1 and RPB2 clades. While this placement does not support the gene duplication model proposed, neither does it rule out that possibility. Because of much higher substitution rates in RPD1/RPE1 as compared to RPB1 (20 fold) and in RPD2 as compared to RPB2 (10 fold) (Table 3.3), as well as their novel functions, the apparent phylogenetic positions of RNAP IV subunits may reflect differing evolutionary rates and altered mechanistic constraints, rather than evolutionary history.

Intron position comparisons provide an alternative means of documenting gene duplication events. The 13 intron positions of RPB1 and the 24 intron positions of RPB2 genes are conserved in Zygnematales, Coleochaetales, Charales, and land plants. RPD1 and RPE1 are identical to one another in their intron positions, five of which are also shared with the N-terminal region of *Arabidopsis* and *Spirogyra* RPB1 genes. This strongly suggests that RPD1/E1 was duplicated from RPB1 through a genomic DNA duplication. On the other hand, there are seven introns in the RPD2 genes, none of which coincides with any of the 24 RPB2 intron positions common to land plants and green algae, nor with those of *A. thaliana* RPA2 (23 positions) or *A. thaliana* RPC2 (36 positions). This suggests that RPD2 was duplicated by a process that did not conserve introns (i.e. by reverse transcription). These considerations, together with our observations that RPD1 exists in Charales but RPD2 does not, indicate that RNAP IV came into being by a multistep process (Fig 3.4A). Initially, the largest subunit of RNAP II underwent duplication to form RPD1; after the divergence of land plants from charalian algae, the RPD2 gene arose. Before this second gene duplication, the largest subunit of RNAP II and the newly duplicated largest subunit RNAP IV would have shared a common second largest subunit. A parallel case exists for the more recent duplication of RPD1/RPE1 in angiosperms. There are two largest subunit genes (RPD1 and RPE1), but only one second-largest subunit gene (RPD2) in angiosperms. The sequences of RPD1 and RPE1 are highly divergent from each other, being 25% identical and 40% similar at the protein level in both *Arabidopsis* and *Oryza*. In this case, two divergent largest subunits appear to share a common second-largest subunit in forming

two different enzymes with nonredundant functions (Kanno et al., 2005). Just as RPB1 gave rise to RPD1, the RPD1/RPE1 duplication may represent the initial step toward evolution of an additional RNA polymerase.

3.4.2 The model for RNAP I, II and III evolution:

The three nuclear RNA polymerase (RNAPs) I, II and III are present in the nuclei of all eukaryotic organisms thus far exmined. It is not known, however, how these three RNAPs evolved from a presumed single prokaryotic ancestor or how they acquired their specialized roles in transcription. Evolutionary homology of major subunit sequences and overall subunit content links the three nuclear RNA polymerases of eukaryotes to the single RNA polymerase of Archaea (Langer et al., 1995). Indeed, several proposals for eukaryotic biogenesis envision a eubacterial-archaeal fusion with the genes encoding RNA polymerases coming from the archaeal partner(Puhler et al., 1989; Golding and Gupta, 1995). The mode of subsequent evolutionary change to give three differentiated systems for nuclear transcription, namely RNAP I (rRNA), RNAP II (pre-mRNA) and RNAP III (tRNA and 5SrRNA) has neither been described nor hypothesized. The process most reasonably began by gene duplication events, since the archeal two largest RNAP subunits share many homologous domains with the corresponding RNAP I, II and III subunits. For all multisubunit RNA polymerases, the combined structure of these two proteins encompasses all the major catalytic surfaces of the enzyme.

We present here a model for the early events in eukaryotic RNA polymerase diversification (subfunctionalization; Force et al., 1999) in eukaryotes, events so ancient as to confound extrapolations of their nature. By analogy with the gene duplication event that gave rise to RNAP IV, we propose (Fig. 3.4B) that a similar multistep mechanism operated in the early evolutionary divergence of RNAPs I, II and III. Initially, gene duplications gave rise to the three largest subunits of RNAPs I, II and III. Subsequently, duplication of the second-largest subunit gene gave rise to RPB2 and a common second largest-subunit gene for RNAP I/III. Finally, an additional duplication gave rise to RPC2

and RPA2. Our phylogenetic analyses of the largest subunit genes suggest that RPA1 diverged first and that RPB1 and RPC1 are sister taxa. While this agrees with the conclusions of earlier studies (Puhler et al., 1989; Zillig et al., 1989), support for the branches in question is not strong, suggesting a nearly simultaneous divergence of the three largest subunit families. On the other hand, RPA2 and RPC2 are sister to each other with strong support (ML = 96%, PP = 100%) (Fig. 3.2), implying that the duplication that produced these genes occurred a reasonable length of time after the divergence of RPB2 from the ancestral second largest subunit gene. Further support for a shared evolutionary history of RNAP I and III comes from RPAC40 and RPAC19, homologs of the alpha subunit of eubacterial RNAP, which are shared between RNAPs I and III in *S. cerevisiae* and in all other eukaryotes (Dequard-Chablat et al., 1991; Ulmasov et al., 1995; Dacks et al., 2002; Goff et al., 2002; Hu et al., 2002; Yu et al., 2002). These two subunits interact with each other and with the second-largest subunit to form an intermediate in enzyme assembly (Kimura et al., 1997; Cramer et al., 2001).

Our multi-step model for the evolution of RNAPs I-III (Fig. 3.4) suggests that divergence of the largest subunit is the first step in the evolution of a new RNAP and that initially, the diverged largest subunits share a common second largest subunit. This feature of the model could be tested by determining whether RPD1 physically interacts with RPB2 in the Charales and whether hybrid RNA polymerases with non-cognate largest and second-largest subunits can be formed. Domain substitution and gene disruption experiments in yeast could be used to make such constructs. Our model (Fig. 3.4) envisions a process whereby both major RNAP subunits can be transformed into daughter molecules capable of heterologous dimerization, while still preserving the parental enzyme. One important aspect of generating new RNA polymerase molecules, namely descent with modification, is thereby accounted for, leaving open questions relating to RNA polymerase specialization. How, for example, did RNAP I acquire the ability to transcribe a single RNA product (40S pre-rRNA) rapidly and processively and how did the other two nuclear RNA polymerases develop their specialized capabilities of rapid reinitiation (RNAP III) and adaptability to a wide variety of DNA templates (RNAP

II)? Perhaps detailed structural comparisons of RNAP IV with RNAP II will provide additional insights into these questions.

Table 3.1 List of taxa and genes studied.

Higher taxon	Species	Genbank #
Glaucocystophyte (Glaucocystaceae)	*Glaucocystis nostochinearum*	RPB1: DQ202658
Green algae		
Chlamydomonadales(Chlamydomonadaceae)	*Chlamydomonas reinhardtii*	RPB2: DQ020659
Zygnematales (Zygnemataceae)	*Spirogyra sp. UWCC FW 670*	RPB2: DQ029103
Coleochaetales(Colecochaetaceae)	*Colecochaete scutata*	RPB2: DQ029105
Charales (Characeae)	*Chara australis*	RPD1: DQ020646; RPB2: DQ029104
	Chara hispida	RPD1: DQ020645
	Nitella clavata	RPD1: DQ020647
	Nitella hyalina	RPD1: DQ020648
Marchantiophyta (Liverworts)		
Marchantiaceae	*Marchantia polymorphia*	RPD1: DQ020651
Lunulariaceae	*Lunularia cruciata L.*	RPD2: DQ011644
Anthocerotophyta (hornworts)		
Anthocerotaceae	*Anthoceros sp. NCU-3e*	RPD1: DQ020650
Bryophyta (mosses)		
Sphagnaceae	*Spagnum sp. UWBC*	RPD1: DQ020652; RPD2: DQ011648 RPB2: DQ029106
Lycopodiophyta (club mosses)		
Selaginellaceae	*Selaginella erythropus.*	RPD2: DQ011647 RPB2: DQ029107
Lycopodiaceae	*Lycopodium sp. UWBC1571*	RPD2: DQ011645
Isoetaceae	*Isoetes sp. UWBC1963*	RPD2: DQ011641
Psilotophyta		
Psilotaceae	*Psilotum nudum*	RPD2: DQ011646 RPB2: DQ029108
Filicophyta (ferns)		
Pteridaceae	*Ceratopteris thalictroides*	RPD1: DQ020649
Spermatophyta (seed plants)		
Ginkgophyta (Ginkgoaceae)	*Ginkgo biloba*	RPD2: DQ011643
Poales (Poaceae)	*Zea mays*	RPD2: DQ000671
Brassicales (Brassicaceae)	*Arabidopsis thaliana*	RPE1: DQ020656, RPD1: DQ020657; RPD2b: DQ029109
Ericales (Ericaceae)	*Rhododendron macrophyllum*	RPD2: DQ011640
Lamiales (Plantaginaceae)	*Antirrhinum majus*	RPD1: DQ 020655; RPD2: DQ011642
Solanales (Solanaceae)	*Solanum lycopersicum*	RPD1: DQ020654; RPE1: DQ020653; RPD2: DQ011649

Table 3.2 The genes coding for the largest and second-largest subunits of RNA polymerases in eukaryotes. The protein genbank accession numbers are indicated. The sequences were gotten through BLAST search against each genome whose sequencing has been completed or is near completion.

Sources	RNAP I		RNAP II		RNAP III	
	RPA1	RPA2	RPB1	RPB2	RPC1	RPC2
Fungi						
Encephalitozoon cuniculi GB-M1	NP_584825	NP_597555	NP_597540	NP_586140	NP_585937	NP_586343
Cryptococcus neoformans var. neoformans JEC21	XP_570516	XP_571380	XP_570943	XP_570204	XP_571468	XP_572718
Schizosaccharomyces pombe 972h-	NP_596300	NP_595819	NP_595673	Q02061	NP_595506	NP_593690
Debaryomyces hansenii CBS767	XP_458147	XP_461338	XP_456921	AAT12540	XP_462102	XP_462552
Yarrowia lipolytica CLIB99	XP_505388	XP_503752	XP_501909	XP_502376	XP_502142	XP_500966
Kluyveromyces lactis NRRL Y-1140	XP_456115	XP_451816	XP_455310	XP_451784	XP_454912	XP_455128
Eremothecium gossypii	NP_984470	NP_982975	NP_984182	NP_985951	NP_985109	NP_983821
Candida glabrata CBS138	XP_445928	XP_447785	XP_447415	XP_448959	XP_449275	XP_448895
Saccharomyces cerevisiae	NP_014986	NP_015335	NP_010141	NP_014794	NP_014759	NP_014850
Aspergillus nidulans FGSC A4	*	EAA59242	EAA65639	EAA61953	EAA63997	EAA65727
Neurospora crassa	EAA26770	EAA35588	EAA34861	EAA27870	EAA27335	EAA27959
Gibberella zeae PH-1	EAA73617	EAA75198	EAA67568	EAA69938	EAA67744	EAA70436
Metazoa						
Caenorhabditis elegans	NP_496872	NP_492476	NP_500523	NP_498047	NP_501127	NP_498192
Drosophila melanogaster	NP_523743	NP_476708	NP_511124	NP_476706	NP_573071	NP_523706
Mus musculus	NP_033114	NP_033112	NP_033115	NP_722493	XP_487324	NP_081699
Homo sapiens	NP_056240	NP_061887	NP_000928	NP_000929	NP_008986	NP_060552
Chlorophyta						
Chlamydomonas reinhardtii	C_380107	C_570093	C_1140054	DQ020659 (C_50164)	C_6040001	C_70231, C_70232
Apicomplexa						
Plasmodium falciparum 3D7	NP_703439	NP_701218	NP_473294	NP_473071	NP_705181	NP_701431
Theileria annulata	CAI73289	CAI76377	CAI75406	CAI76382	CAI76663	CAI75960
Cryptosporidium hominis	EAL35354 EAL35355	EAL38145	EAL35927	EAL38006	EAL37032	EAL34724
Cryptosporidium parvum	*	EAK88354	EAK89262	EAK90367	EAK89955	EAK87469
Entamoebidae						
Entamoeba histolytica HM-1:IMSS	EAL49525	EAL46961	EAL46395	EAL48102	EAL48362	EAL49502
Dictyosteliida						
Dictyostelium discoideum	EAL69630	EAL60592	EAL67745	EAL63310	EAL68785	EAL63250

Table 3.2 (continued)

Diplomonadida						
Giardia lamblia ATCC 50803	EAA40851	EAA39886	EAA41730	EAA38052	EAA37804 +EAA37805	EAA42548
Kinetoplastida						
Trypanosoma brucei	P16355	CAD26832	AAX80548(A); P17545(B)	AAX80202	CAA31014	CAB95348
Cryptomonad nucleomorph						
Guillardia theta	CAC27032	AAK39710	CAC27103	AAK39815	AAK39916	AAK39687

*: unfinished genome project

Table 3.3 The genes coding for the largest and the second-largest subunits of RNA polymerases in *Arabidopsis* and *Oryza* nuclear genome.

		pol I	pol II	pol III	pol IV	
Largest subunit β' homolog	Arabidopsis	RPA1 (At3g57660)	RPB1(At4g35800)	RPC1 (At5g60040)	RPD1 (At1g630200)	RPE1(At2g40030+At2g 40040)
	Oryza	RPA1(BAD35511) (AACV01011176 + AACV01014535+ AACV01014536)	RPB1(XP_493925) (AACV01010889, AACV01017298, AACV01017300)*	RPC1(XP_472999) (AACV01010028 +AACV01010029)	RPD1a (AACV01010330) RPD1b (AACV01020526+ AACV01020527) I: 88% S:94%	RPE1a (AACV01003372) RPE1b (AACV01003092) I:42%, S: 58%
	Substitution rate (d)	0.348	0.049	0.245	1.034/1.02	0.965/0.543
Second-largest subunit β homolog	Arabidopsis	RPA2(At1g29940)	RPB2(At4g21710)	RPC2(At5g45140)	RPD2 (a:At3g18090, b:At3g23780) I: 96%, S: 98%	
	Oryza	RPA2(NP922143) (AACV01021863)	RPB2(XP480298) (AACV01007472)	RPC2(XP_470900) (AACV01006851)	RPD2(XP_474064) (a: AACV01010557, b: AACV01017377)\ I: 70%, S: 80%.	
	Substitution rate (d)	0.543	0.053	0.261	0.448/0.585	

Gene loci in Arabidopsis are indicated. The contigs and genebank access number for protein sequences are shown only for *Oryza sativa* ssp. *japonia*. "+" indicates the same gene in different contigs. * There are three copies of RPB1 genes in the Oryza genome. The DNA sequences of AACV01017298 and AACV01017300 are 99% identical to each other and 93% identical to DNA sequences of AACV01010889. Only AA sequences are used for the calculation of sequence homology. I: Identity; S; Similarity. The relative rate (d) of evolution, calculated by the number of substitution per site, between Arabidopsis and Oryza orthologs are calculated only from the conserved regions that are used for phylogeny in this paper.

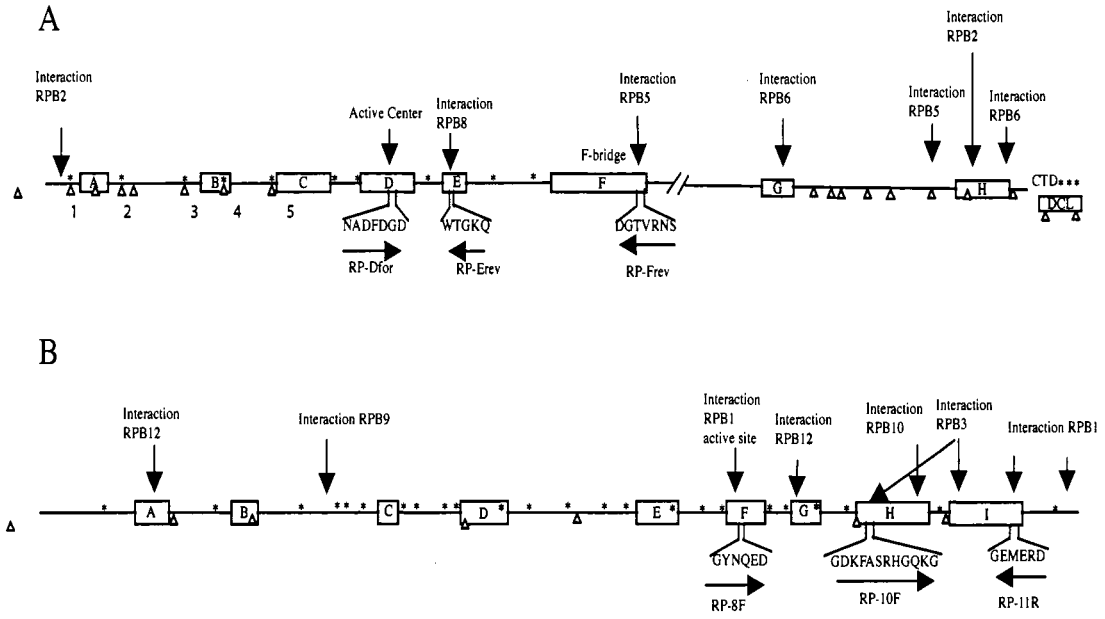


Figure 3.1. The Structure of largest (A) and second-largest (B) subunits of RNA polymerases. The conserved regions in the largest and second-largest subunits have been shown in rectangles. The stars above the lines show the intron positions in RPB1 and RPB2 of Arabidopsis thaliana. The triangles below the lines show the the intron position of RPD1/E1 and RPD2 in A. thaliana. The interaction sites are based on RNAP II 3-D structure by Cramer et al., 2002. RPD1 and RPE1 have completely deleted sequences between region F and G. The DCL domain (Bellaoui and Gruissem, 2004) exists at the C-terminal of RPD1/RPE1 genes in Angiosperms but not in bryophytes.
The primers are shown as arrows.

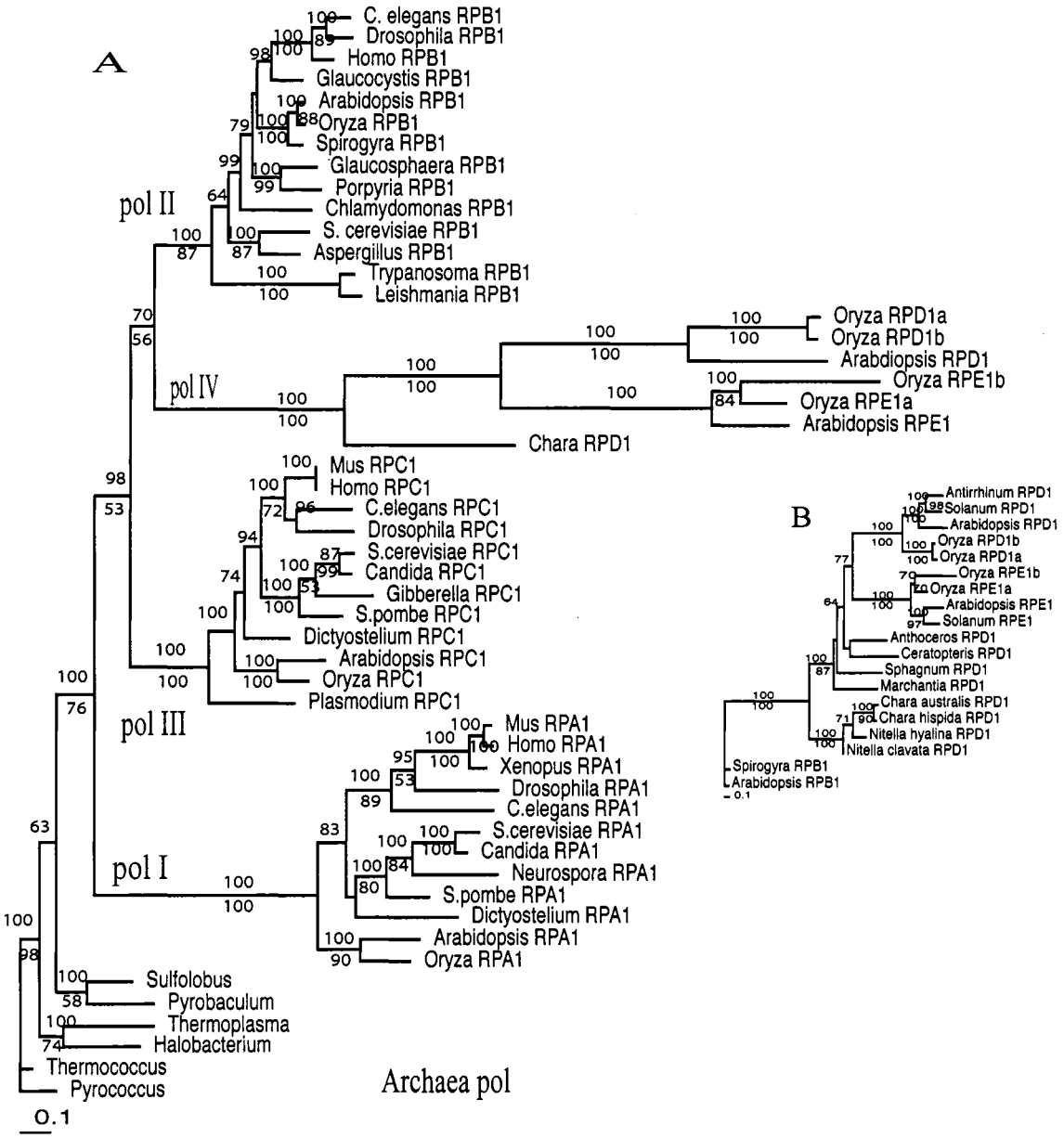


Figure 3.2 A. The phylogenetic relationships of the largest RNAP subunit genes in eukaryotes and archaea using protein sequences of conserved regions A through H. B. The RPD1/E1 ML gene tree using sequences between regions D and H. ML bootstrap values >50% are shown below the branches and the posterior probability of Baysian inference above the branches.

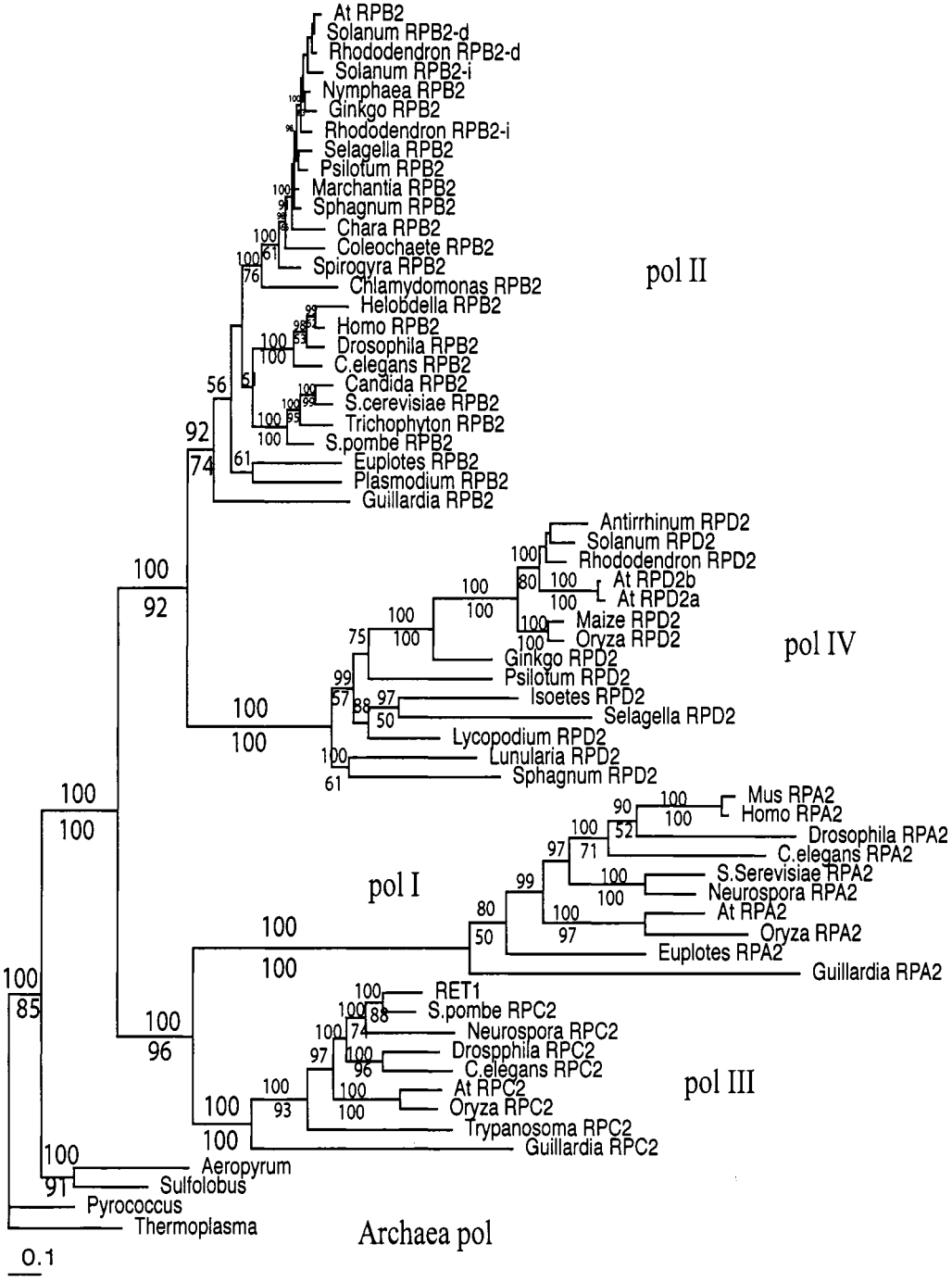


Figure 3.3 The phylogenetic relationships of the second-largest subunits of RNAPs in eukayotes and archaea. ML and Bayesian support numbers are as in Fig. 3.2.

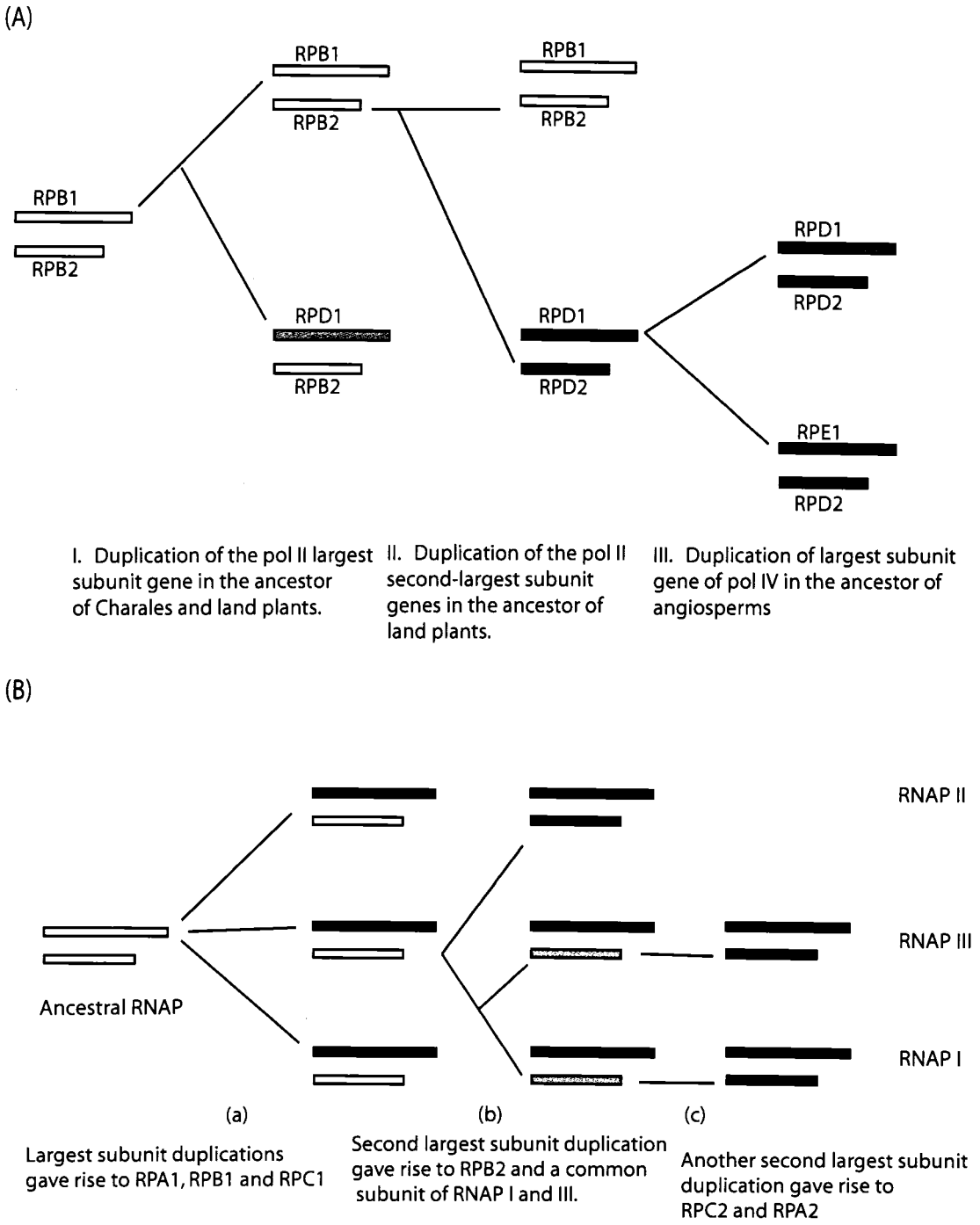


Figure 3.4 A. The model for the evolution of RNAP IV from RNAP II in plants. B. The model for the evolutionary divergence of eukaryotic RNAPs. The long and short bars represent the largest and second-largest subunits, respectively. For the RNAP I, II and III products of evolution, both subunits are the same color (green, red and blue, respectively). For evolutionary intermediates, sharing of the same second-largest subunit between different polymerases is indicated by white or brown.

Chapter 4. Characterizing RNA polymerase IV complexes in cauliflower (*Brassica oleracea var. botrytis*).

Summary

I have found genes for the second-largest subunit (RPD2) of a new DNA-dependent RNA polymerase, pol IV, in all land plants and genes for the largest subunit (RPD1/RPE1) of this enzyme in all major land plant taxa and in closely related green algae. Here I report an immunochemical and chromatographic study of this enzyme in cauliflower (*Brassica oleracea var. botrytis*). There are two pol IV protein holozymes: an RPD1/RPD2 complex and an RPE1/RPD2 complex. The RPE1/RPD2 enzyme is the major form found in cauliflower inflorescence. In chromatographic separations on phosphocellulose columns, the RPE1/RPD2 complex is co-purified with pol Ia and pol IIIa activity peaks. Through tandem mass spectrometry, I found that pol I and pol III have distinct third largest subunits (RPAC40a and RPAC40c, respectively) in cauliflower. A different third largest subunit (RPB3a, b) is shared by pol II and pol IV. Based on our results, we propose that pol IV is a multi-subunit enzyme that contains a β' homolog (RPD1 or RPE1), a β homolog (RPD2), α homologs (RPB3a, b and RPB11), and common subunits RPB5, RPB6, RPB8, RPB10 and RPB12. The largest and second largest subunits of pol IV have all the conserved domains and sites for the active center, NTP binding, and DNA/RNA hybrid binding. Dramatic changes at sites contacting the incoming template DNA suggest to us that pol IV is a polymerase either using single stranded DNA as template or one with low processivity. Pol IV is non-essential for Arabidopsis growth and development and Pol IV mutants are defective in asymmetric DNA methylation.

Key words: RNA polymerase IV, RPD1, RPE1, RPD2, Brassica oleracea var. botrytis

4.1 Introduction

In plants, the three RNA polymerases have been purified from wheat germ (Jendrisak and Burgess, 1975) and cauliflower inflorescence (Guilfoyle et al., 1976). Cauliflower (*Brassica oleracea var. botrytis*), like *Arabidopsis thaliana* a member of the family Brassicaceae, is ideal for purification of RNA polymerases. As in other systems, pol I and pol III can be separated from pol II polymerase activity by DEAE-Cellulose chromatography. They are eluted as a single peak of RNA polymerase activity from DEAE-Cellulose chromatography. Pol I and pol III proteins can be separated by subsequent DEAE-Sephadex chromatography. In cauliflower, pol I can be further separated into two peaks of RNA polymerase activity by phosphocellulose chromatography. These are named pol Ia and pol Ib (Guilfoyle et al., 1976). The subunit compositions of RNA polymerases in plants are similar to those in yeasts, animals and protists (Guilfoyle and Jendrisak, 1978). In *Arabidopsis*, the genes encoding most of the pol II subunits have been cloned. In contrast to other organisms, *Arabidopsis* has two copies of *RPB3* (Larkin and Guilfoyle, 1996; Ulmasov et al., 1996), and six copies of *RPB5* (Larkin et al., 1999). Both copies of *RPB3* and at least two copies of *RPB5* (*RPB5*a and *RPB5* b) are expressed in *Arabidopsis* (Larkin et al., 1999). The detailed molecular physiology of these paralogous subunits remain to be determined. It was suggested that different copies of *RPB5* may be associated with different transcription factors in regulating specific gene transcription (Miyao and Woychik, 1998; Larkin et al., 1999).

There are three and only three sets of genes coding for the largest (RPA1, RPB1, RPC1) and second-largest (RPA2, RPB2, RPC2) nuclear RNAP subunits for fungi, animals and protists. We have found genes (Chapter 3) for the largest subunit (RPD1/RPE1) of a new DNA-dependent RNA polymerase, Pol IV, in all major land plant taxa and in closely related green algae and genes for the second-largest subunit (RPD2) of this enzyme in all land plants. Pol IV evolved from pol II through a multiple step mechanism in the ancestors of Charales and land plants. There are two genes for the

largest subunit of pol IV, RPD1 and RPE1, in angiosperms. These are the products of a gene duplication that occurred before angiosperm divergence (Chapter 3). We report here our study of the protein properties of pol IV in cauliflower and *Arabidopsis*.

4.2 Material and Methods

4.2.1 Protein extraction and ion-exchange chromatography

Cauliflower inflorescences (*Brassica oleracea var. botrytis*) were obtained in the supermarket. All purification procedures were carried out at 4 °C. Buffer A: 25 mM MES-NaOH (pH 6.0), 20 mM MgCl2, 20mM KCl, 0.25M Sucrose, 10mM 2-mercaptoethanol, 40% Glycerol. Buffer B: 0.05M Tris-HCl (pH 8.0), 0.1mM EDTA, 10mM DTT, 0.5 mM PMSF, 10-30% Glycerol. The DEAE-Cellulose chromatography, DEAE-Sephadex chromatography, and phosphocellulose chromatography were prepared as described by Jendrisak and Burgess (1975) (Jendrisak and Burgess, 1975).

Nuclear protein extraction followed the procedures of Guilfoyle (1976) with slight modifications. Cauliflower tissue (about 5 Kg) homogenized in buffer A (about 4 L) at low speed to keep the nuclei intact. The homogenized tissue was filtered through 8 layers of cheesecloth and the pass-through was centrifuged at 5500 rpm (Beckman, rotor JA-10, ~5,000g) for 30 mins. The pellets were washed with buffer A + 1% Triton X-100. The RNA polymerase was solubilized from the nuclei by suspending the pellets in 40 ml buffer B (10% glycerol) + 0.5 M ammonium sulfate, shearing the pellets for 2 min at high speed with a Polytron and sonicating the chromatin for 3 min on ice (Ultrasonics Inc, W-140-D, 20 watts). The solubilized RNA polymerase was recovered in the supernatant after centrifugation at 50,000rpm (Beckman, rotor Ti70) for 1 hour. The RNA polymerase was precipitated by addition of solid ammonium sulfate (0.33g/mL) and stirred for 1 hour. The protein precipitate was recovered by centrifugation for 1 hour at 11,000 rpm (Beckman, rotor JA-20, ~15,000g).

DEAE-cellulose chromatography. The ammonium sulfate precipitate was suspended in buffer B and adjusted to 0.05 M ammonium sulfate. The solution was loaded onto a column (25 X 1.5 cm) of DEAE-Cellulose equilibrated with buffer B + 0.05 M ammonium sulfate. After the column was washed with 1 column volume (~ 40 ml) of buffer B + 0.05M ammonium sulfate, the enzymes were eluted with a 0.05 M to 0.3 M linear gradient of ammonium sulfate in buffer B. The gradient was 100 ml for 0.25 M of ammonium sulfate and 2-ml fractions were collected.

DEAE-Sephadex chromatography. Fractions under each peak of RNA polymerase after DEAE-Cellulose chromatography were combined and diluted to 0.05 M ammonium sulfate with buffer B. Then peak fraction was chromatographed on a DEAE-sephadex column (20 X 1.5 cm) with a linear gradient from 0.05 to 0.4 M ammonium sulfate. The gradient was set 70 ml for 0.35 M ammonium sulfate and 2-ml fractions were collected.

Phosphocellulose chromatography. Fractions under each peak of DEAE-chromatography were combined and diluted to 0.05 M ammonium sulfate, loaded onto phosphocellulose column (12 X1.5 cm). For pol I and III, the enzymes were eluted at a gradient of 70 ml for 0.35 M ammonium sulfate from 0.05 M to 0.4 M ammonium sulfate. For pol II, the enzymes were eluted at a gradient of 100 ml for 0.25 M ammonium sulfate from 0.05 to 0.3 M ammonium sulfate.

4.2.2 RNA polymerase activity assay

RNA polymerase activity was assayed in a reaction volume of 100 ul by combining the following: 50 ul reaction buffer (0.1 M Tris-HCl (pH 8), 50% glycerol, 10 mM $MgCl_2$, 50mM $(NH_4)_2SO_4$), 5 ul single stranded DNA template (Sigma D9156), 5 ul each of mM unlabelled ATP, CTP and GTP, 20-30 ul enzyme from each fraction and 0.5 to 1 ul 3H-UTP. The reaction was incubated on a PCR thermocycler at 28 °C for 30 min. 90 ul of each reaction was transferred onto Whatman GF/A filter discs (24 mm in

diameter) and the filters were immediately immersed in ice cold 10% TCA to terminate the reaction and stirred for 10 mins. Then the filters were washed three time in 5% TCA three times and 95% ethanol two times and left for air-drying. The radioactivity is then determined by liquid scintillation counting.

4.2.3 Immunoprecipitation and western blotting analysis

Antibodies. The cDNA of *A. thaliana* RPD2 gene between the Hind III and Xho I cutting sites (position 969 to 1975) was cloned into pET-28a expression plasmid (Novagen). The protein was expressed in Tuner (DE3) cells upon IPTG induction and purified by His-Bind purification kits (Novagen, #70239-3). Then the purified proteins were injected to rabbit to generate antibodies. For peptide antibodies, the peptide sequence for RPE1 is RPE1-1: SESAINHPSQLINAC; the peptide sequence for RPD1 is RPD1-1: SPSSNTKVPLSPWVC; the peptide sequence for RPD2 is IKSTKFPPAESVDEC. The rabbit peptide antibodies were ordered from Synpep Corporation. The antibodies against specific peptide were then immuno-purified by passing the antiserum though the gel coupled with the peptides (Pierce, #20401).

Immunoprecipitation. Protein extraction was carried out as described above. After ammonium sulfate precipitation, the pellets were suspended in buffer B. The solution was then loaded onto a Sephadex G-50 column (10 X 1.5 cm) equilibrated with PBS buffer and then eluted with PBS buffer. The protein fractions in PBS buffer were incubated with immuno-purified antibodies or prebleed serum for 3 hours on ice. Then agarose-protein A was added into the solution and incubated for one hour with constant rolling at 4 °C. The agarose-protein A was collected by centrifugation and washed three times with IPP buffer: 50 mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 0.1% DTT, 0.1% Triton X-100, 20% Glycerol, 150 mM NaCl and once with IPP buffer + 200 mM NaCl. The enzymes were eluted by incubation for 1 hour with IPP buffer + 200 mM NaCl + 0.1 M peptides. The elution was repeated twice and the combined eluted proteins were concentrated by microfiltration using centricon YM-100 (Millipore).

Electrophoresis and Western-blotting. The 7.5% gel in 0.25M Tric-HCl (pH 8.8) SDS-PAGE is used for resolving gels (for 10 ml gel: 1.9 ml 40% acrylamide/Bis 37.5:1 (Ambion #9026), 5.37 ml H_2O, 2.5 ml 1M Tris-Hcl ph 8.8, 100ul 10% SDS, 100ul 1% peroxydisulphate and 20 ul TEMED) and 4.5% gel in 0.125M Tric-HCl (pH6.8) is used for stacking gel (for 5ml gel: 550ul 40% acrylamide/Bis 37.5:1, 3.75 ml H2O, 625ul 1M Tris-Hcl ph 6.8, 50ul 10% SDS, 50 ul 1% peroxydisulphate and 10 ul TEMED). After electrophoresis, the proteins were electro-blotted to PVDF membrane at 200 mA for 1 hour by Genie Blotter (Idea Scientific Co.). The transfer buffer is 48mM Tris-39mM Glycine (pH 9.2) with 0.01% SDS. The Vectastain ABC-AmP kit (Vector Laboratories, Inc.) is used for western blotting.

4.2.4 Mass-spectrometry

The total protein amount in each RNA polymerase activity peak is about 20 ug. First, the proteins were precipitated by 20% TCA on ice for 20 mins (by adding 1 volume of 100% TCA to 4 volumes of the protein samples) and spin at 14 K rpm in microcentrifuge for 10 mins at 4 °C. The protein pellets were washed twice with 200 ul cold acetone and dried in the vacuum drier. The Mass spectrometry analyses were carried out by Dr. Claire Delahunty in Dr. John Yates's lab at The Scripps Research Institute, CA. The MS/MS spectra were searched against the Arabidopsis protein database to identify the polypeptides.

4.2.5 Arabidopsis mutants and bisulfite genomic DNA sequencing.

All the *Arabidopsis thaliana* mutants were obtained from seed stock of The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org). The seed stocks for the pol IV gene mutants are: SALK_083051 for RPD1, SALK_029919 for RPE1, SALK_095689 for RPD2b, SALK_008535 for RPD2a (Fig. 4.3A). The T-DNA insertions have been confirmed by PCR and sequencing with Lba1 primer on the T-DNA and a flanking reverse primer on the genes where the T-DNA is inserted. The primers

used for amplify wild-type and insertion mutant alleles are: Lba1: atggttcacgtagtgggccatcg ; RPD1-051F: gcactcggttttggaactgg ; RPD1-051R: gagactatcagctacgagaagc; RPE1-GF2: gatagactcatgtctttcagtgc; RPE1-HR: tgagcctgtaccaacatccacgc ; MRC-ex1F: ttgacgagatcgagtctgctggg ; RPD2a-ex3R: atctgatgttcaacatatgttagg ; RPD2b-in6F: tgttgattcttgagtcctgattca ; RPD2b-3'R: accaaaccaaaacctatactacta .

Bisulfite genomic DNA sequencing. The bisulfite DNA sequencing procedure I used is adapted from a protocol in Steven Jacobsen's lab: 1). digest 4ug genomic DNA with restriction enzymes (EcoRI and Hind III) for 6 hours and purified by Qiagen PCR purification kits. 2). Take 40 ul purified DNA and heat at 97 °C for 5 min and put immediately on ice. 3). Add 2u l 6.3M NaOH (5.04g/20 ml) and incubate at 39 °C for 30 min. 4). Add 426 ul freshly prepared bisulfite solution [dissolve 4.05g sodium bisulfite to 8 ml H20, adjust to pH 5.1 with NaOH and add 330ul 20mM hydroquinine (1.1g hydroquinone/500 ml), adjust volume to 10 ml] and incubate at 55 °C in the PCR thermocycler for 18 hours, punctuated every three hours by a 5 minute denaturation step at 95 °C. 5). Desalt the samples by Microcon YM-30 (Millipore) and wash the column three times. 6). Wash the column with 0.3 M NaOH twice and incubate the column at RT for 20 min between the washes. 7). Wash the column three times with water and twice with 0.1 M TE buffer. 8). Collect the DNA and perform two rounds of PCR. 9). The fresh PCR product is cloned into TA vector (Invitrogen) and at least 14 clones were checked for each genotype. The primers for the At-SN1 sites are: SN1-F1: ggtcacacagyggtagaaataatgtttyttggg ; SN1-F2: aaattaaayaaaataagtggtggtgta ; SN1-R1: tatractcctcctcaacaaaaataattcc ; SN1-R2: aataatactttractaactttcractccc .

4.3 Results

4.3.1 RNA polymerase chromatography

Both anion-exchange (DEAE-Cellulose and DEAE-Sephadex) chromatography and cation-exchange (phosphocellulose) chromatography was used to purify nuclear

RNA polymerases from cauliflower inflorescence. More than twenty enzyme preparations were made by the procedures described below. The purification procedures are robust and repeatable.

The nuclear extract was first analyzed by DEAE-Cellulose chromatography. RNA polymerase activity elutes as two overlapping peaks at 0.1 M and 0.18 M ammonium sulfate (Fig. 4.1A). The first peak is resistant to alpha-amanitin and can be further separated by DEAE-Sephadex chromatography into two peaks of RNA polymerase activity eluting at 0.15 M and 0.27 M ammonium sulfate (Fig. 4.1B). These two RNA polymerases are designated as the pol I activity peak and the pol III activity peak. The pol I activity can be further resolved as two partially overlapping peaks of RNA polymerase activity on phosphocellulose chromatography at 0.13 M and 0.20 M ammonium sulfate (Fig. 4.1D). These two peaks have been designated as pol Ia and pol Ib (Guilfoyle et al., 1976). The pol III activity chromatographed as a single peak at 0.20 M ammonium sulfate on phosphocellulose (Fig. 4.1F). The second peak from DEAE-Cellulose chromatographed as a single peak at 0.20 M ammonium sulfate on DEAE-Sephadex. At 5ug/ml, α-amanitin totally inhibits pol II polymerase activity. While the majority of RNA polymerase activity eluted from this column is sensitive 5 ug/ml α-amanitin, a small fraction of the enzyme activity is α-amanitin resistant (Fig. 4.1C). On subsequent phosphocellulose chromatography, this RNA polymerase activity can be further resolved into two peaks of RNA polymerase activity eluting at 0.10 M and 0.13 M ammonium sulfate (Fig. 4.1E). The first peak is α-amanitin sensitive and is designated as pol II. I found later that the second peak has pol III activity. Thus it is designated as pol IIIa and the pol III eluted at 0.20 M ammonium sulfate on phosphocellulose chromatography is designated as pol IIIb. The pol IIIa activity has not been previously reported.

4.3.2 Western blotting and MS-MS

In order to study the chromatographic properties of pol IV, an anti-RPD2 antibody was used to check for the presence of RPD2 protein in different column

fractions. The RPD2 protein is present in both peaks of RNA polymerase activity on the initial DEAE-Cellulose column (Fig 4.1 A) (Fig. 4.2A, lanes 1 and 2). On DEAE-Sephadex chromatography (Fig 4.1B), the RPD2 protein is present in the pol I activity peak but not in the pol IIIb peak (Fig. 4.2A, lanes 3 and 4). On subsequent phosphocellulose chromatographyof the pol I activity peak (Fig 4.1D), the RPD2 protein is mainly in the pol Ia peak and to a slight extent in the pol Ib peak (Fig. 4.2A, lanes 5, 6, and 7). For the pol II + pol IIIa peak, on DEAE-Sephadex chromatography (Fig 4.1C), the RPD2 protein exists in the pol II peak (Fig. 4.1B, lanes 8 and 9). On phosphocellulose chromatography (Fig 4.1E), the RPD2 protein is mainly in the pol IIIa peak and partially in the pol II peak (Fig. 4.1B, lanes 10, 11 and 12).

Tandem mass spectrometry (MS/MS) was used to identify the polypeptide sequences present in each polymerase activity peak. Tandem MS/MS is a powerful way to identify the protein composition of a complex mixture. Proteins in each polymerase activity peak were precipitated by TCA, digested with trypsin, separated by reverse phase-liquid chromatography (RP-uLC) and analysed by electrospray ionization (ESI) MS and tandem mass spectrometry (MS/MS). The MS/MS spectra, generated by the fragmented peptides in the gas-phase by collision-induced dissociation (CID), were used by the algorithm SEQUEST to search against the *Arabidopsis* protein database to identify the protein. The polypeptides encoded by RNA polymerase genes in each polymerase activity peaks are summarized in Table 4.1. For the pol Ia and pol Ib peaks, peptides present were those corresponding to the largest and second largest subunits of pol I, RPA1 and RPA2, the third largest subunit RPAC40a, and the common small subunits RPB5 and RPB8. For the pol II peak, we can identify most of the subunits of the pol II apoenzyme, including those encoded by RPB1, RPB2, RPB3a,b, RPB5, RPB7, RPB8, RPB11, and RPB12. For the pol IIIa and pol IIIb peaks, we can identify peptides corresponding to RPC1, RPC2, RPAC40c, RPB5, RPB6 and RPB8, and a homolog of the 62kD protein of human pol III. RPB5 and RPB8 polypeptides were found in all fractions of pols I, II and III peaks. There are two RPB8 genes in *Arabidopsis thaliana*: at1g54250 and at3g59600. Only at3g59600 has been detected in all three RNA polymerases. There

are six RPB5genes in *A. thaliana*; of these, only RPB5-at3g22320 has been found in all three RNA polymerases. Two peptides corresponding to RPB5-at5g57980 were present in the pol IIIa fraction. There are two genes for the third largest subunit of pol II, RPB3a (at2g15430) and RPB3b (at2g15400). Both of these gene products are present in the pol II fraction. The pol I and pol III fractions have different third largest subunits, AC40a and AC40c respectively. There is some contamination by chloroplast encoded RNA polymerase subunits in the pol II and pol IIIb fractions.

Peptides corresponding to pol IV subunits are found mainly in the pol Ia and pol IIIa activity peaks. Both RPE1 and RPD2 proteins are found there. The gene products of RPB3a and RPB3b are also present in these enzyme peaks. In the pol Ia and pol Ib peaks, no pol II subunit polypeptides were present, suggesting that the presence of RPB3 product cannot be contributed by the pol II complex. Therefore, the presence of the RPB3a and RPB3b gene products implies that they are in this fraction because it contains the pol IV complex. In the pol IIIa fractions, the amount of RPB3a and RPB3b corresponds to the amount of RPE1 and RPD2 detected in this peak. This is further suggestive evidence for the presence of RPB3a and RPB3b in the same complex with RPE1 and RPD2, since a slight contamination by pol II could not contribute all the RPB3a and RPB3b detected in this peak. Even though both RPE1 and RPD1 genes are expressed in the cauliflower inflorescence (data not shown), no RPD1 protein product was detected in any portion tested.

4.3.3 Co-Immunoprecipitation

To test whether RPE1 and RPD2 physically interact with each other and to look for RPD1 proteins, I performed immunoprecipitation using antibodies against the specific peptides of RPE1 and RPD1. The immuno-precipitated protein was then checked by western blotting using antibodies against the RPD2 protein. RPD2 protein can be detected in the immunoprecipitates with both anti-RPE1 and anti-RPD1 but not in those with pre-bleeding sera controls (Fig 4.3A). This suggests that there are two pol IV

complexes: an RPD1/RPD2 complex and an RPE1/RPD2 complex. We checked the RNA polymerase activity of the immuno-purified products. There is no RNA polymerase activity associated with the immuno-purified products, even though we can detect the presence of RPD2 proteins by western-blotting in these samples (Fig 4.3B).

4.3.4 Bisulfite genomic DNA sequencing

In order to test for an *in vivo* function of pol IV, we obtained T-DNA insertion lines in the pol IV genes of *A. thaliana*. The T-DNA insertion lines were crossed to generate homozygous single and double T-DNA insertion mutants of the pol IV genes. These mutants are designated as dd for *RPD1* mutant, ee for *RPE1* mutant, ddee for *RPD1* and *RPE1* double mutant, aa for *RPD2a* mutant, bb for *RPD2b* mutant, aabb for *RPD2a* and *RPD2b* double mutant. These mutant strains have been confirmed to be homozygous by PCR analysis (Fig. 4.4). Each of these homozygous T-DNA insertion mutants was shown to lack function of the gene; no corresponding mRNA can be detected by the RT-PCR in the mutants. All single and double mutants of the largest and the second largest subunits of each of these pol IV genes grows normally and show no obvious phenotypic defect in growth and development. This suggests that pol IV is non-essential for *Arabidopsis* growth and development.

Concurrently with this research, several labs used forward genetics to screen *A. thaliana* mutants for defects in the pathways of si-RNA mediated DNA methylation and gene silencing. They found that knockout mutations of both the largest and second largest subunits of pol IV have a defect in silencing and demethylation of endogenous retroelements (such as At-SN1) and heterochromatic 5S RNA gene repeats (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005). One way to detect the DNA methylation change is to use the technique of bisulfite genomic DNA sequencing. Bisulfite treatment of DNA converts unmethylated cytosine residues into uracil while leaving the methylated cytosine unchanged. After PCR amplification and sequencing, the methylated cytosine

residue will be replaced by cytosine while unmethylated cytosine residue will changed to thymine residue.

Using bisulfite genomic DNA sequencing, we detected the change of DNA methylation at a retroelement site (At-SN1 site) in the pol IV gene mutants (Fig. 4.5). For symmetric CG methylation, no methylation changes occurred in single RPD1 (dd) or single RPE1 (ee) mutants. CNG methylation was not changed in single dd mutants but is reduced in ee mutants. However, asymmetric (CHH) cytosine methylation was abolished in both dd and ee mutants. For the homozygous double mutant of RPD1 and RPE1 mutant (ddee), CG methylation was reduced and both CNG and CHH methylation are abolished. Plants homozygous for mutations in both paralogs encoding the second-largest subunit (aabb) had a defect in DNA methylation like those doubly mutated in genes for the largest subunit (ddee). So RPD1 and RPE1 have partial compensatory function for CNG methylation, and both are essential for asymmetric methylation at the At-SN1 site.

4.4 Discussion

4.4.1 Pol IV subunit composition inferred from chromatography and mass spectrometric analyses

Using protein sequences from yeast and humans, we performed homology searching against *Arabidopsis* genome sequences and identified all the genes coding for subunits of RNA polymerase core enzymes (table 4.2). Through the combination of ion-exchange chromatography and tandem mass spectrometry, we have partially purified the RNA polymerases and identified the subunit composition of nuclear RNA polymerases in cauliflower inflorescence. The Arabidopsis gene products identified by our MS/MS results are summarized in table 4.2. Our mass spectrometry results found that pol I and pol III have different alpha homologs, AC40a and AC40c, respectively, in cauliflower. We have sequenced the cauliflower genes for both the AC40a and AC40c subunits. The molecular weights of their translational products are 40 and 43 kDa. Multiple genes code

PB5, RPB6, RPB8 and RPB12 in *Arabidopsis*, and many orthologs are also present in Brassica (data incomplete). However, for each of these subunits, only a single gene product has been found in one or all RNA polymerase fractions in our MS analysis. This indicates that they are the major proteins used as common subunits for nuclear RNA polymerases in cauliflower. RPB5-at5g57980 has been found in the pol IIIa fraction and antibody against RPB5-3g16680 has detected its protein form in tissue culture of *A. thaliana* (Larkin et al., 1999). Therefore, it is likely that other minor forms of the common subunits are also functional and incorporated into the core enzymes, but their roles in regulating gene transcription, if any, have yet to be determined.

The RPB3a and RPB3b proteins have been found both in fractions that contain pol II and those contain pol IV, suggesting that pol II and pol IV share a common alpha homolog. This finding agrees with our phylogenetic conclusion that pol IV evolved from pol II (Chapter 3). The two alpha homologs in pol II, RPB3 and RPB11, form a heterodimer and are important for pol II assembly. The RPB3 and RPB11 homologs in Arabidopsis have been found to form a heterodimer *in vitro* and *in vivo* (Ulmasov et al., 1996). So it may as well be that RPB3 and RPB11 also form a heterodimer for pol IV assembly.

By mass spectrometry we found the RPE1/RPD2 complex in the pol Ia and pol IIIa peaks. We did not find the RPD1/RPD2 complex in any column fraction even though the co-immunoprecipitation experiment showed that anti-RPD1 peptide can pull down RPD2 proteins. There are several possible explanations for this result: 1, we might have missed it because mass spectrometry was performed only on peak fractions of RNA polymerase activity. The RPD1/RPD2 might differ in chromatographic properties such that it would be eluted in a region with no polymerase activity. 2, the RPD1 protein might have eluded detection by mass spectrometry simply because of low abundance.

4.4.2 Pol IV subunits inferred from protein homologies

The common subunits RPB5, RPB6, RPB8, RPB10 and RPB12 all interact with either the largest or second largest subunits in the yeast pol II structure (Cramer et al., 2001). For RPB5, this interaction is with the "cleft" domain of the largest subunit. The two major interaction sites between RPB5 and RPB1 are conserved in the largest subunits of pol I, pol II and pol III. They are also conserved in the sequences of the largest subunits of pol IV (Fig. 4.6 A and Fig. 4.10). The RPB8 subunit interacts with RPB1 around conserved region E. The same interacting sites are conserved in the largest subunits of pol IV as in those of pols I, II and III, except that there are variable short insertions in several pol IV sequences (Fig 4.6 C). The RPB6 subunit interacts with residues in conserved regions G and H of RPB1. Conserved region G spans over the joint of "Jaw" and "Cleft" domains. It can be separated into two functional motifs: the first motif in the jaw interacts with RPB6; the second motif forming a G-loop interacts with the bridge helix at the active center. Only the sites that interact with the RPB6 are conserved in the largest subunit of pol IV (Fig 4.6 B(left); Fig. 4.10). In conserved region H, the first several residues are conserved in the largest subunit of pol IV, while the rest of residues preserve similar residue properties with rest of polymerases (Fig 4.6 B, right).

Both the RPB10 and RPB12 subunits interact with residues in conserved regions in the second largest subunits of pol II (RPB2). The RPB10 subunit interacts with residues in conserved regions F and H. RPB12 subunits interacts with residues in conserved regions A and G. All these sequences are also conserved in the second largest subunit of pol IV (Fig 4.7).

Not only are these common subunits shared among pol I, pol II and pol III; in addition the yeast RPB6, RPB8, RPB10, and RPB12 subunits can be functionally substituted by their human homologs (McKune et al., 1995; Shpakovski et al., 1995), suggesting that the common subunits have a broad range of interaction capacity. Since

the sites interacting with common subunits are also conserved in the largest and second largest subunits of pol IV, we think it is reasonable to infer that these common subunits are also incorporated into the pol IV protein complex. So we conclude, tentatively, that pol IV is also a multiple subunit enzyme like other nuclear RNA polymerases, including a largest subunit (RPD1 or RPE1), a second largest subunit (RPD2), alpha homologs RPB3 and RPB11, and common subunits RPB5, RPB6, RPB8, RPB10 and RPB12.

4.4.3 The pol IV RNA polymerase activity

The RPE1/RPD2 complex identified by mass spectrometry co-purified with pol Ia and pol IIIa. Further purification of pol IV protein failed due to the small amount of protein, precluding a determination of whether the RPE1/RPD2 complex has RNA polymerase activity. Immunopurified pol IV also shows no RNA polymerase activity, even though we can detect the RPD2 protein in the immuno-purified products. On the face of it, these data suggest that pol IV has no RNA polymerase activity. However, we have found that our traditional RNA polymerase assay requires at least 100 ng of protein. Dilution of the protein amount 100 times will reduce the level of assayable polymerase activity to the background level. The amount of pol IV protein is estimated to be 100 to 1000 times less than that of pol II protein. Therefore, it is likely that the failure to detect RNA polymerase activity associated with immuno-purified pol IV results from the small amount of protein present. At this stage, we cannot conclude experimentally whether pol IV has RNA polymerase activity or not.

4.4.4 Bioinformatic inference of an RNA polymerase activity of pol IV

Regarding Pol IV function in plants, its absence does not affect plant survival and in fact it is not needed for normal plant growth and development. The higher substitution rates of the Pol IV genes in plants may be explained by the non-essential nature of the enzyme. It is also possible pol IV may have fewer interacting partner proteins, allowing rapid evolution of pol IV genes.

To find out whether the inferred polypeptide sequences of the two largest RNAP IV subunits are consistent with enzymatic activity, we compared them to those of the RNAPs I, II and III. The sequences around the active center, DNA/RNA hybrid binding, and the incoming NTP binding sites are conserved in all multiple-subunit RNA polymerases (Kettenberger et al., 2004; Westover et al., 2004). The corresponding sites for active center, DNA/RNA hybrid binding, and NTP binding are also conserved in pol IV (Fig 4.8 and Fig 4.9), supporting its possible role in RNA polymerase activity. Our phylogenetic study shows that the pol IV diverged from pol II by a two-step process. Before duplication of the second largest subunit gene, the largest subunit of pol II and the newly duplicated pol IV largest subunit likely shared a common second largest subunit. The MS/MS evidence and sequence comparison suggest that pol IV and pol II also share the alpha homologs and other common subunits. This would be consistent with an RNA polymerase activity of pol IV.

Suggestive evidence contraindicative of polymerase activity is the presence in the largest subunits of RNAP IV of drastic changes in the region of the F-bridge, the conserved region G and a complete deletion of the "foot" domain in the yeast RNAP II structure (Fig 4.10). The sites lacking in the F-bridge are those that contact and stabilize unwound template DNA at the active center. The G-loop sequences that interact with the F-bridge during RNA polymerase translocation(Epshtein et al., 2002; Vassylyev et al., 2002) also are not conserved in the largest subunit of RNAP IV sequences. These changes in RNAP IV, at the very sites that contact incoming template DNA, suggest to us that, if RNAP IV is a DNA-dependent RNA polymerase, it either uses single stranded DNA as template or has very low processivity.

4.4.5 The pol IV function in Arabidopsis

Unlike pols I, II, and III whose function is essential for cell growth, pol IV is non-essential for normal growth and development in *Arabidopsis*. Evidence for a functional

role for pol IV in *Arabidopsis* comes from studies of the RNAi pathway. Knock-out mutations in RPD1, RPE1 and RPD2 genes have partial defects in siRNA-mediated gene silencing and heterochromatin modification (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005). We studied the methylation changes in the pol IV mutants at one of the retroelement sites (At-SN1), at which asymmetric cytosine methylation is controlled by components in the si-RNA pathway (Hamilton et al., 2002; Zilberman et al., 2003; Chen, 2004). We found that both RPD1 and RPE1 mutations abolish asymmetric DNA methylation at the At-SN1 site, confirming a role of each in si-RNA mediated DNA methylation. In Arabidopsis, asymmetric DNA methylation is carried out by DNA methyltransferase DRM1/DRM2 (Chan et al., 2005). It is not known how pol IV function is integrated into the si-RNA pathway. RPD1 mutations block the generation of si-RNA species, while RPE1 mutations have no effect on the generation of si-RNAs (Herr et al., 2005; Kanno et al., 2005), suggesting that the RPD1 and RPE1 complexes may have different roles in si-RNA gene silencing. Since there is just one largest subunit gene in Charales and bryophytes, the function of pol IV in these taxa, if it has a role in the si-RNA mediated pathway, would include both of these steps.

Small interference RNA (siRNA) has been found to mediate heterochromatin modification in *Schizosaccharomyces pome* (Volpe et al., 2002), *Tetrahymena thermophila* (Yao et al., 2003), *Drosophila melanogaster* (Pal-Bhadra et al., 2004), and *Arabidopsis thaliana* (Zilberman et al., 2003; Lippman and Martienssen, 2004). A point mutation in the RPB2 gene in *S. pombe* uncoupled the RNA polymerase activity from the si-RNA mediated process, suggesting the involvement of pol II in this pathway (Kato et al., 2005). However, it is still unclear how pol II is integrated into the siRNA pathways in *S. pombe* and whether polymerase activity of pol II is needed for siRNA pathway. At this stage, we cannot confirm that pol IV is a DNA dependent RNA polymerase. Understanding whether pol IV has RNA polymerase activity will be central to understanding its role in participation in the si-RNA pathway.

Table 4.1. Summary of MS-MS results. The first numbers are the actual number of peptides found by MS (the total AA residues found, the percentage of the AA residues found)

Samples	Peptides of RNA polymerases been detected by MS-MS		
pol Ia	**pol I** RPA1 15(225, 13%) RPA2 10(129, 10%) AC40-a 7(100, 27%) RPB8-3g 2(41, 28%) RPB5-3g 5(49, 24%)	**pol IV** RPE1 16(215, 11%) RPD2 29(360, 30%) RPB3a,b 7(100, 31%)	
pol Ib	**pol I** RPA1 20(320, 19%) RPA2 15(155, 13%) RPB8-3g 2(41, 28%) RPB5-3g 3(39, 19%) AC40-a 5(82, 22%)	RPE1 2(24, 1%) RPD2 2(32, 3%) RPB3a.b 3(47, 15%)	
pol II	**pol II** RPB1 43(354 , 24%) RPB2 34(297, 25%) RPB3a,b 13(119, 50%) RPB5-3g 4(32, 15%) RPB7 3(38, 21%) RPB8-3g 2(41, 28%) RPB11 3(38, 32%) RPB12-5g 3(16, 31%)	**Chloroplast RNAP** RPOB 6(80, 7.4%) RPOC1 5(53, 9.0%) RPOA 2(30, 8.5%) RPOC2 12(148,10.6%)	 RPE1 4(38, 1.9%) RPD2 6(80, 6.8%)
pol IIIa	**pol III** RPC1 14(188, 14%) RPC2 14(121, 11%) RPC62kD 3(28, 5%) RPB5 -3g 4(41, 29%) RPB8-3g 2(41, 28%) AC40-c 14(194, 50%)	**pol IV** RPE1 15(203, 11%) RPD2 18(278, 24%) RPB3a,b 7(100, 31%)	 RPB1 6(103, 5.5%) RPB2 6(87, 7.3%) RPB11 2(26, 22%) RPB5-5g 2(23, 11%)
pol IIIb	**pol III** RPC1 22(345, 25%) RPC2 17(219, 18%) RPC62kD 3(49, 9%) RPB6-5g 2(53, 37%) RPB5 -3g 6(60, 29%) RPB8-3g 2(41, 28%) AC40-c 9(176, 47%)	**Chloroplast RNAP** RPOB 12(226, 21%) RPOC1 9(138, 20%) RPOA 3(50, 15%) RPOC2 6(94, 7%)	

Table 4.2. Genes encoding subunits of RNA polymerases in Arabidopsis. The polypeptides confirmed by MS analysis or immunoprecipitation are underlined and bolded.

	pol I	pol III	pol II	pol IVa	pol IVb
β'	RPA1 (**at3g57660**)	RPC1 (**at5g60040**)	RPB1 (**at4g35800**)	RPD1 (**at1g63020**)	RPE1 (**at2g40030+40**)
β	RPA2 (**at1g29940**)	RPC2 (**at5g45140**)	RPB2 (**at4g21710**)	RPD2 (**at3g23780**, at3g18090*)	
α	RPAC40a (**at1g60850**) RPAC43b (**at1g60620**)		RPB3a (**at2g15430**) RPB3b (**at2g14500**)		
α	RPAC19 (**at2g29540**)		RPB11 (**at3g52090**)		
common subunit	RPB6 (a: **at5g51940**, b: at2g04630) I: 92%, S: 97% RPB5 (a:**at3g22320**, b:at3g16680, c:**at5g57980**, d:at2g41340, e:at3g57080, f:at3g54490) RPB8 (a: at1g54250, b: **at3g59600**, I: 95%, S: 99%) RPB10 (at1g61700) RPB12 (a: **at5g41010**, b: at1g53690)				
specific subunit	RPA12.2 (at3g25940)	RPC11(at4g07950, at1g01210)	RPB9(at3g16980, at4g16265)		
			RPB4(at5g09920)		
	RPA43(at1g75670)	RPC25(at1g06790)	RPB7(**at5g59180**)		
		RPC62kd (**at3g49000**)			

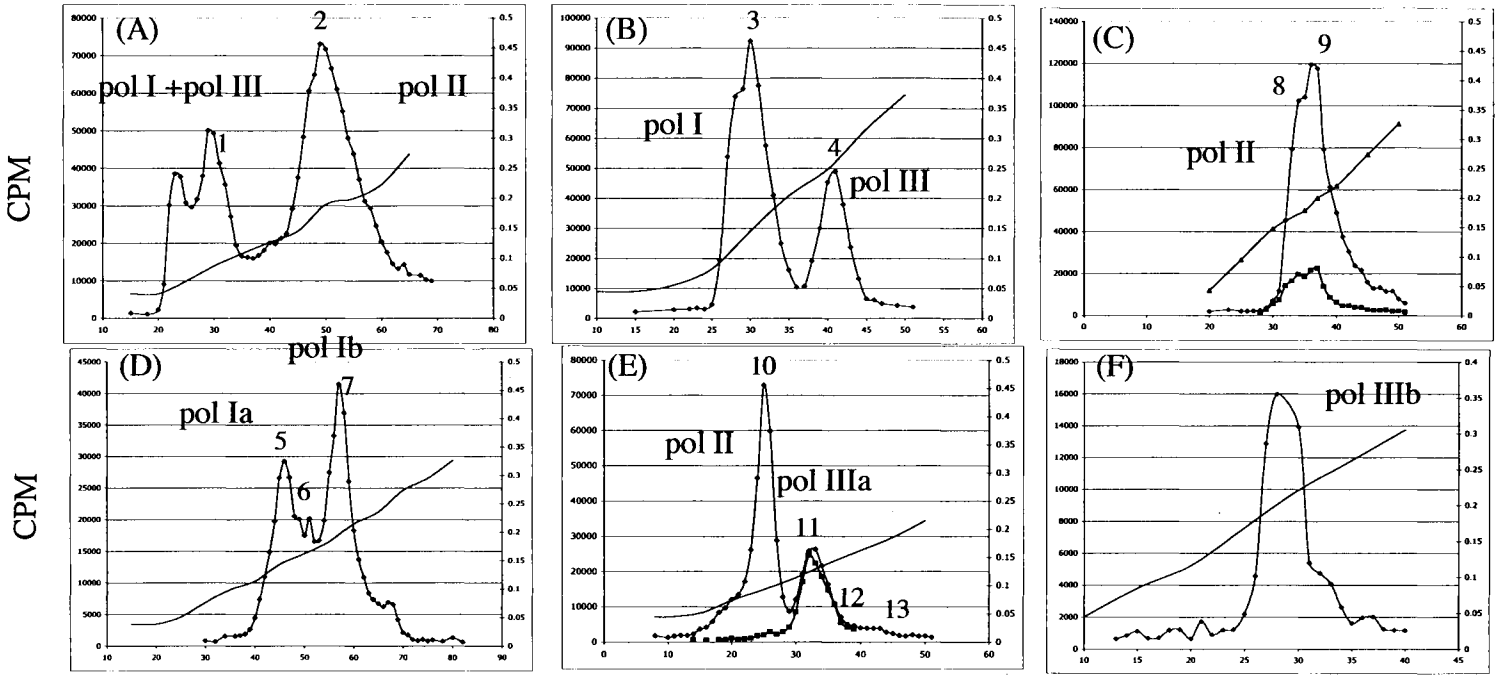


Figure 4.1 The purification of RNA polymerases from cauliflower. (A), DEAE-cellulose chromatography of RNA polymerases. Fractions 18 to 41 (RNA polymerase I + III) and fractions 42 to 70 (RNA polymerase II) were collected. (B), DEAE-Sephadex chromatography of RNA polymerase I and III. (C), DEAE-Sephadex chromatography of RNA polymerase II. ♦---♦: RNA polymerase activity without α-amanitin. ■---■: RNA polymerase activity with 5ug/ml of α-amanitin, (D), Phosophocellulose chromatography of RNA polymerase I from fractions 22 to 36 of (B). (E), Phosphocellulose chromatography of RNA polymerase II from fractions 26 to 54 of (C). ■---■: RNA polymerase activity with 5ug/ml α-amanitin. (F), Phosocellulose chromatography of RNA polymerase III from fractions 37 to 54 of (B).

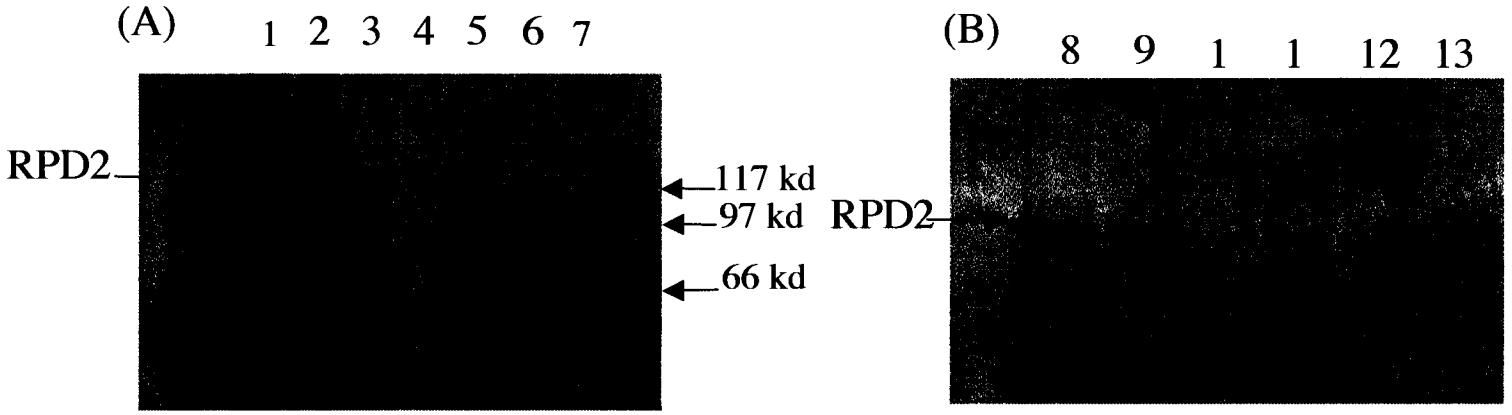


Figure 4.2 (A) and (B) Immuno-blotting using Anti-RPD2 antibody. The RPD2 in *A. thaliana* is about 127 kD. The numbers indicate the samples from peaks of RNA polymerase activity from Fig 4.1 (A) to (E).

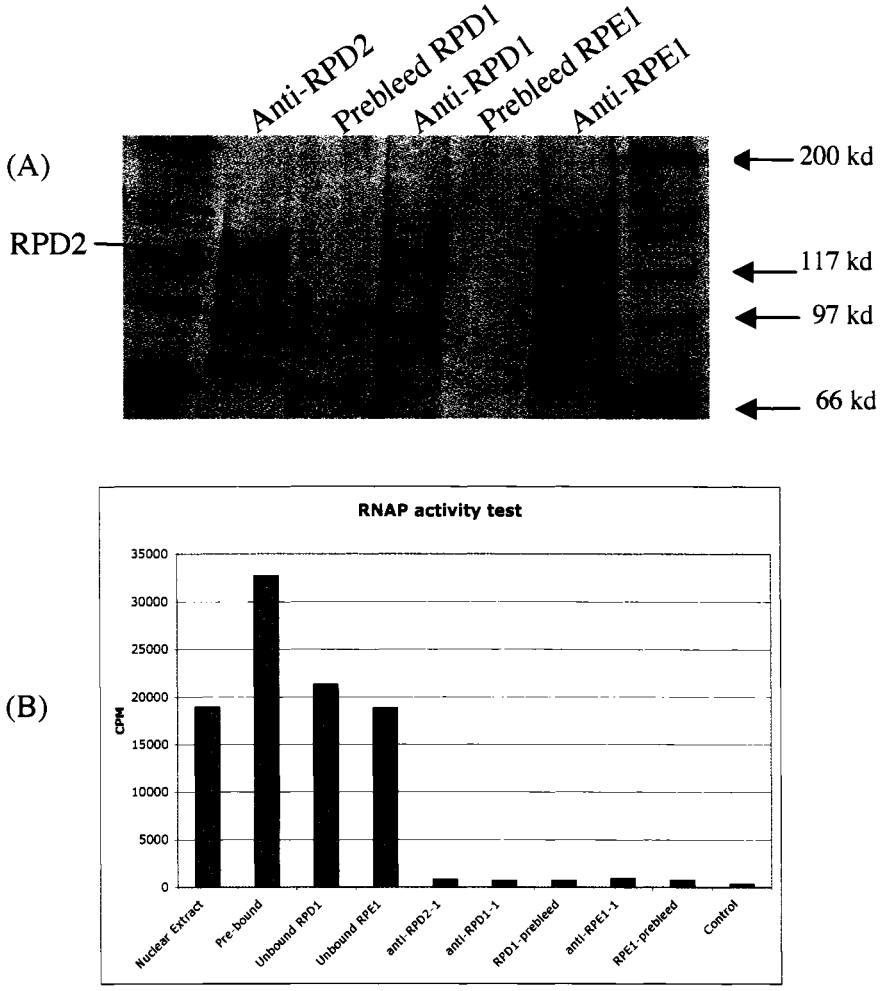
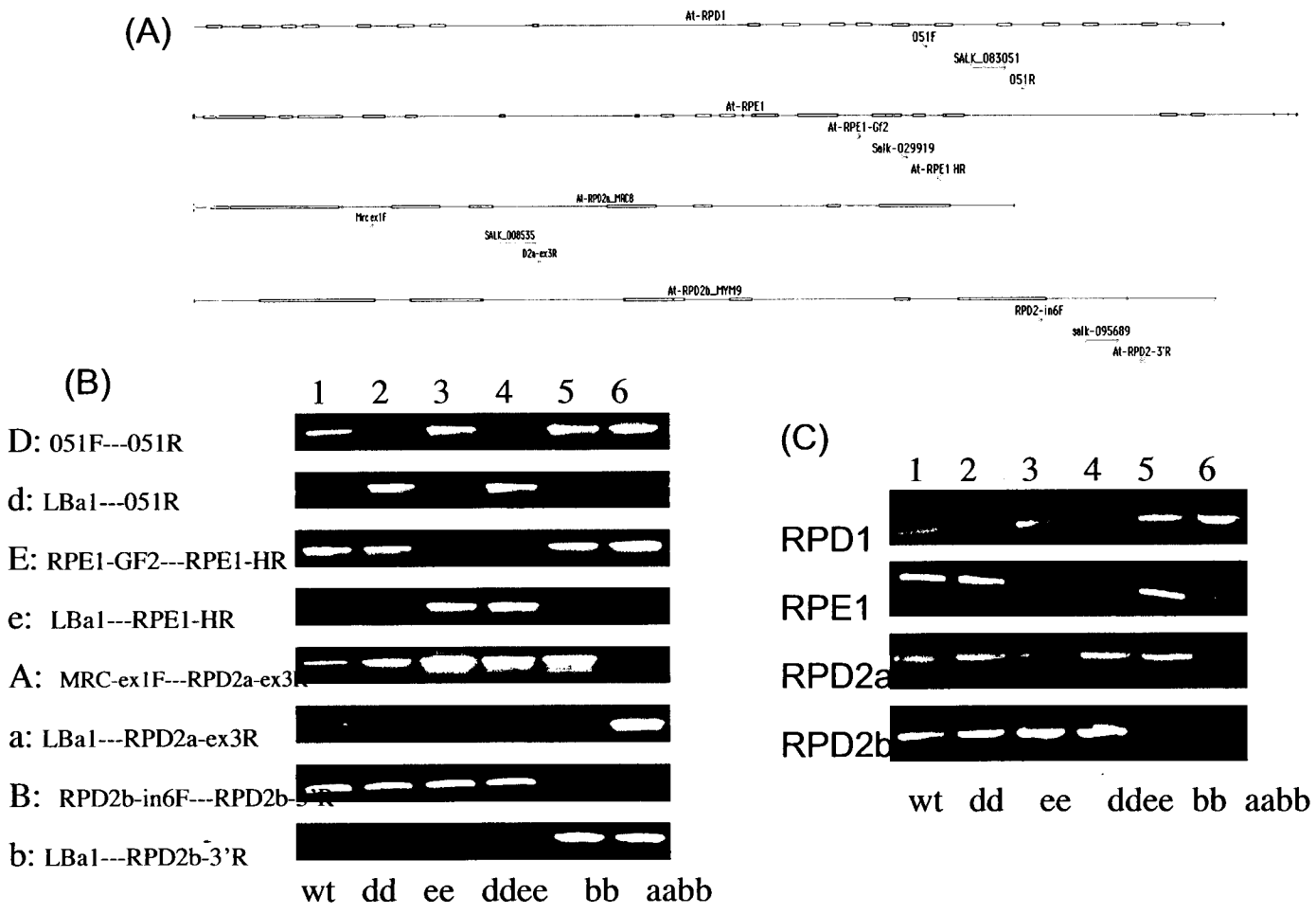
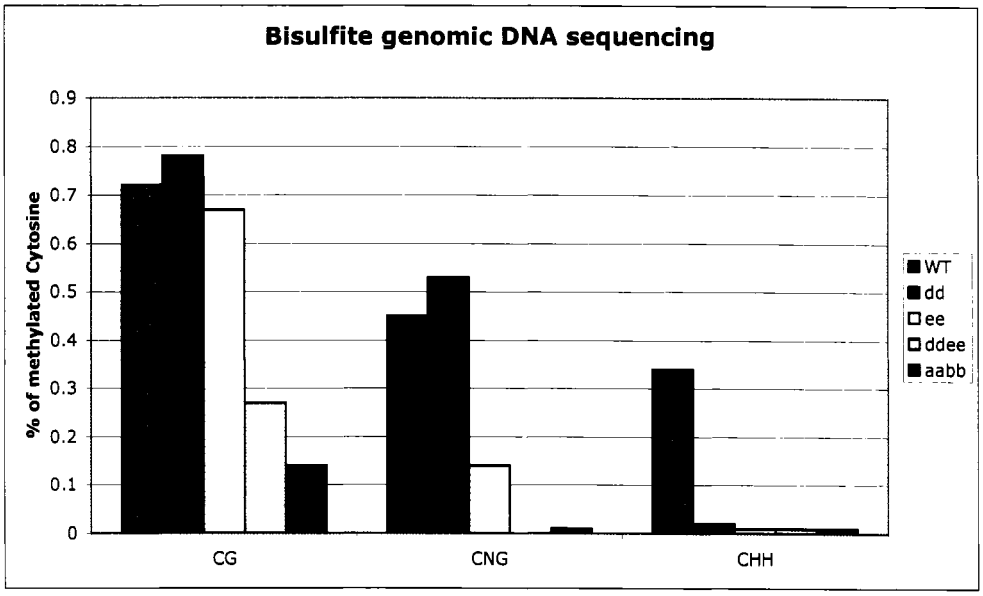


Figure 4.3 Immuno-precipitation of pol IV and RNA polymerase activity analysis. (A), Immuno-blotting using anti-RPD2 antibody. Samples are from immnuo-purified product by peptide antibody or prebleed sera. (B). RNA polymerase activity assay of the unpurified and immunopurified proteins

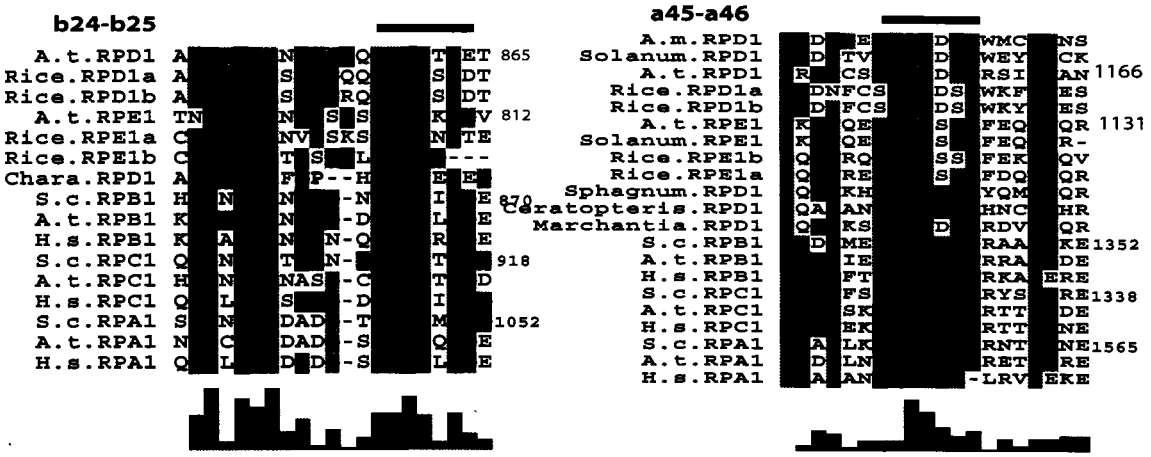

(A)
At-RPD1
051F
SALK_083051
051R
At-RPE1
At-RPE1-Gf2
Salk-029919
At-RPE1 HR
At-RPD2a_MRC8
Mrc ex1F
SALK_008535
D2a-ex3R
At-RPD2b_MYM9
RPD2-in6F
salk-095689
At-RPD2-3'R
(B)
1 2 3 4 5 6
D: 051F---051R
d: LBa1---051R
E: RPE1-GF2---RPE1-HR
e: LBa1---RPE1-HR
A: MRC-ex1F---RPD2a-ex3R
a: LBa1---RPD2a-ex3R
B: RPD2b-in6F---RPD2b-3'R
b: LBa1---RPD2b-3'R
wt dd ee ddee bb aabb
(C)
1 2 3 4 5 6
RPD1
RPE1
RPD2a
RPD2b
wt dd ee ddee bb aabb



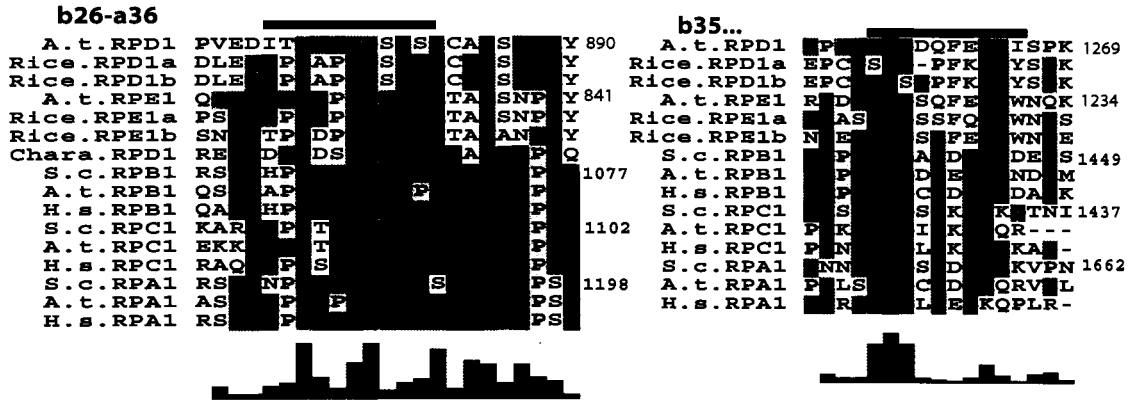
	CG	CNG	CHH
WT	0.72 (52/72)	0.45 (57/126)	0.34 (217/630)
dd	0.78 (53/68)	0.53 (64/119)	0.02 (10/595)
ee	0.67 (38/56)	0.14 (14/98)	0.01 (4/490)
ddee	0.27 (24/88)	0 (0/154)	0.01 (10/770)
aabb	0.14 (8/56)	0.01 (1/98)	0.01 (3/490)

Figure 4.5 DNA methylation at At-SN1 site measured by bisulfite genomic DNA sequencing. The table shows the percentage of methylated cytosine at CG, CNG and CHH positions in different genotypes, which is derived from the actual number of methylated cytosine observed divided by the total number of cytosine in each category, shown in the paresis. The percentage of methylated cytosine in different genotypes is also illustrated in the figure above.

(A) Sites interacting with RPB5



(B) Sites interacting with RPB6



(C) Sites interacting with RPB8

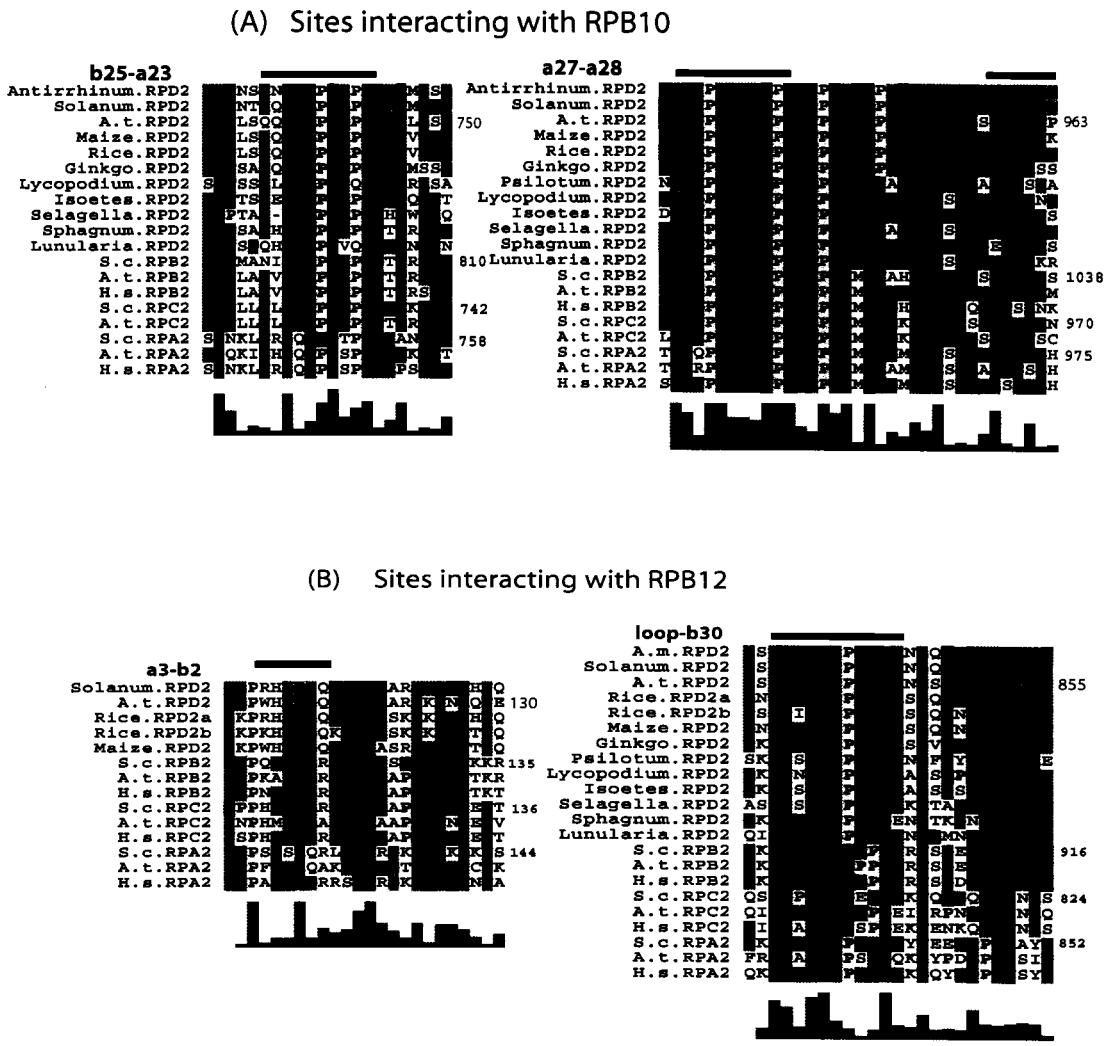


Figure 4.7 The alignment of sites in the second-largest subunits of RNAPs I, II, III and IV that interact with common subsunits of RPB10 and RPB12.
The bar shows the interface in the second-largest subunit of RPB2 that interacts with the small subunits in S. cerevisiae pol II structure.
The color scheme and histograms are used as described in the legend of Fig 4.6.

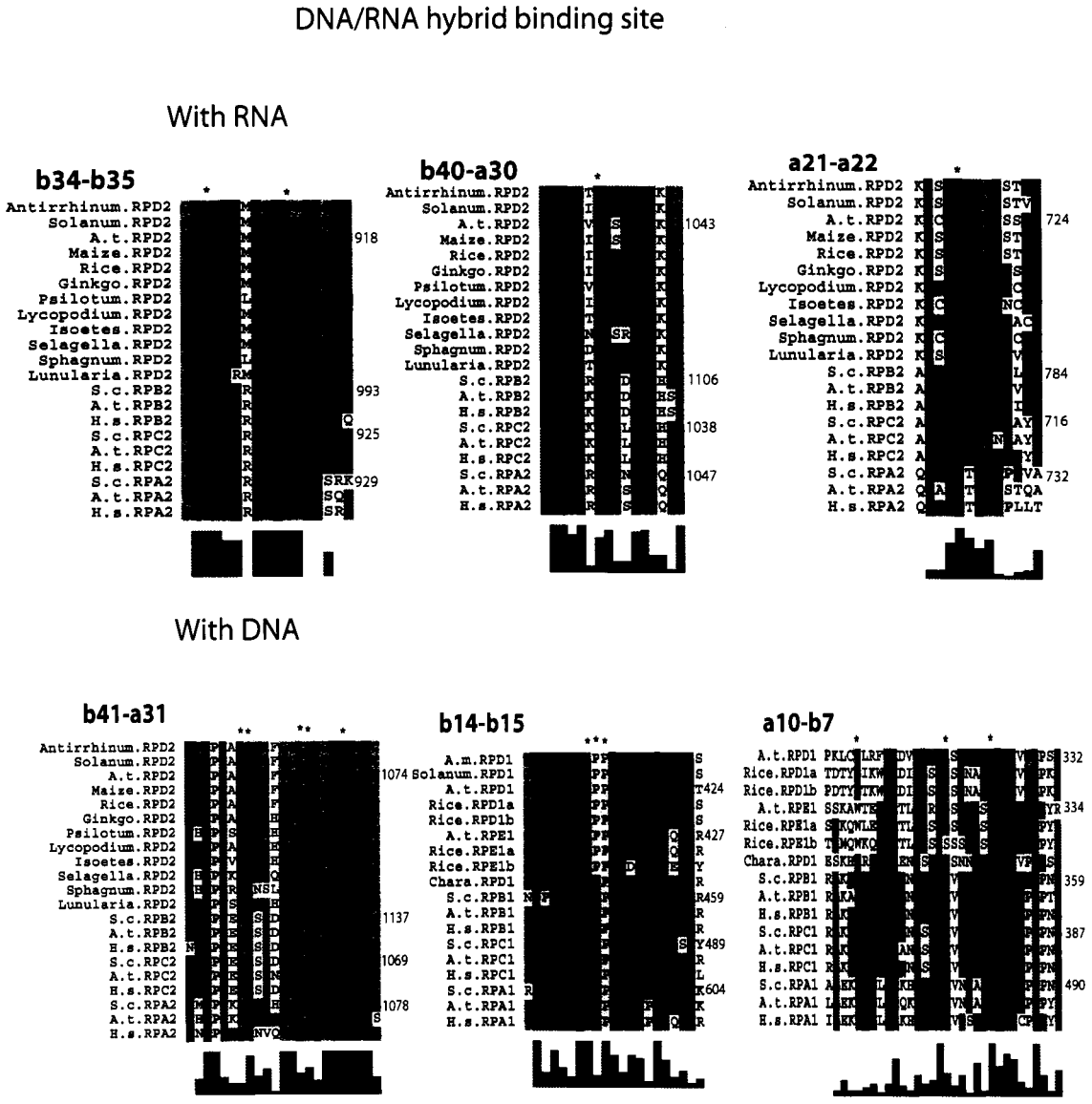


Figure 4.8 The sites in the largest and second largest subunits for DNA/RNA hybrid binding. The hybrid binding sites are derived from crystal structure of yeast pol II elongation complex (Kettenberger et al., 2004). The residues that in close contact with DNA or RNA are marked as *. The color scheme and histograms are used as described in the legend of Fig 4.6

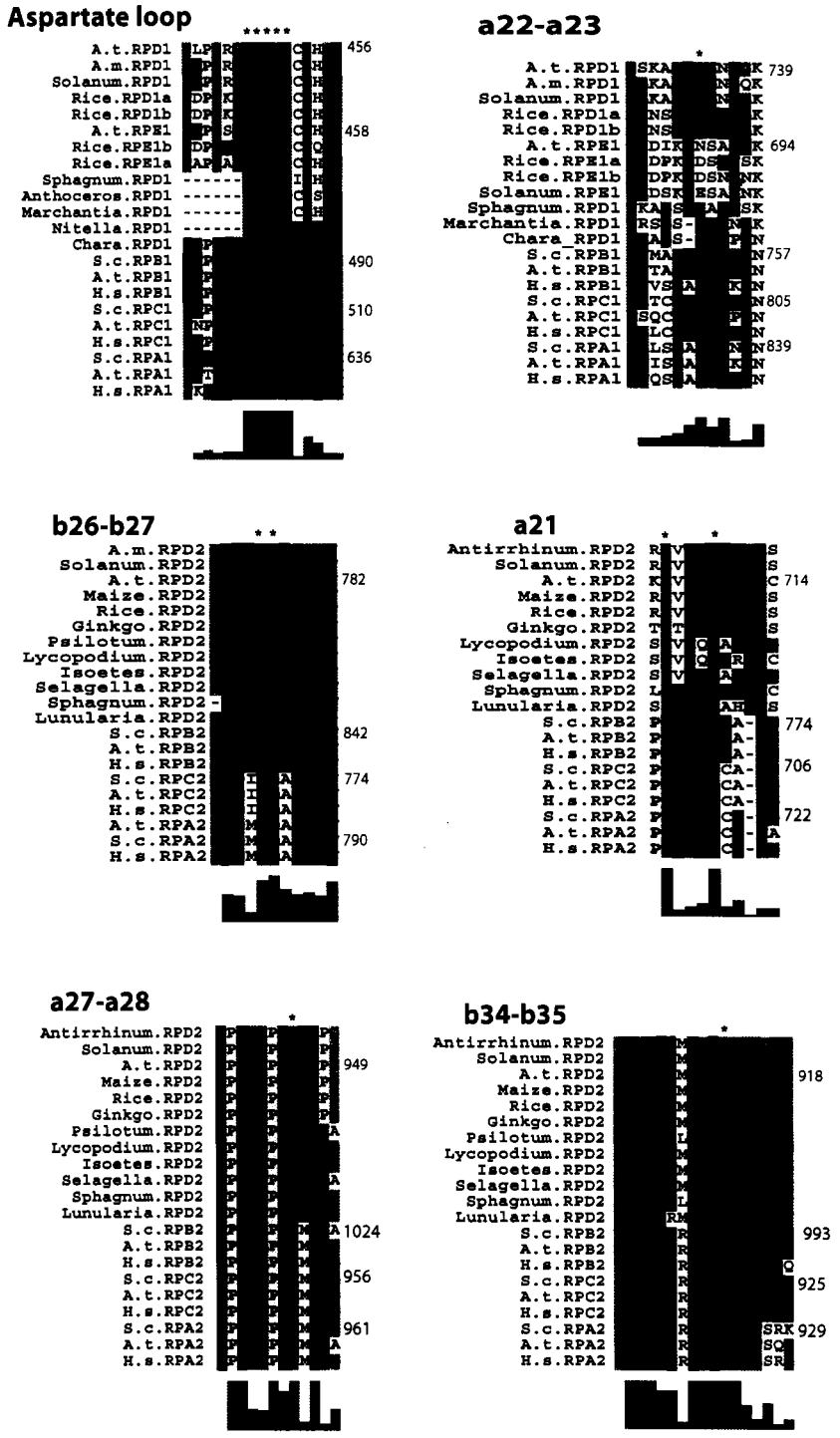


Figure 4.9 The NTP binding sites in the largest and second largest subunits. The NTP binding sites are derived from the crystal structure of pol II elongation complex with NTP (Kettenberger et al., 2004). The residues that are important for NTP binding are marked as *. The color scheme and histograms are used as described in the legend of Fig 4.6.

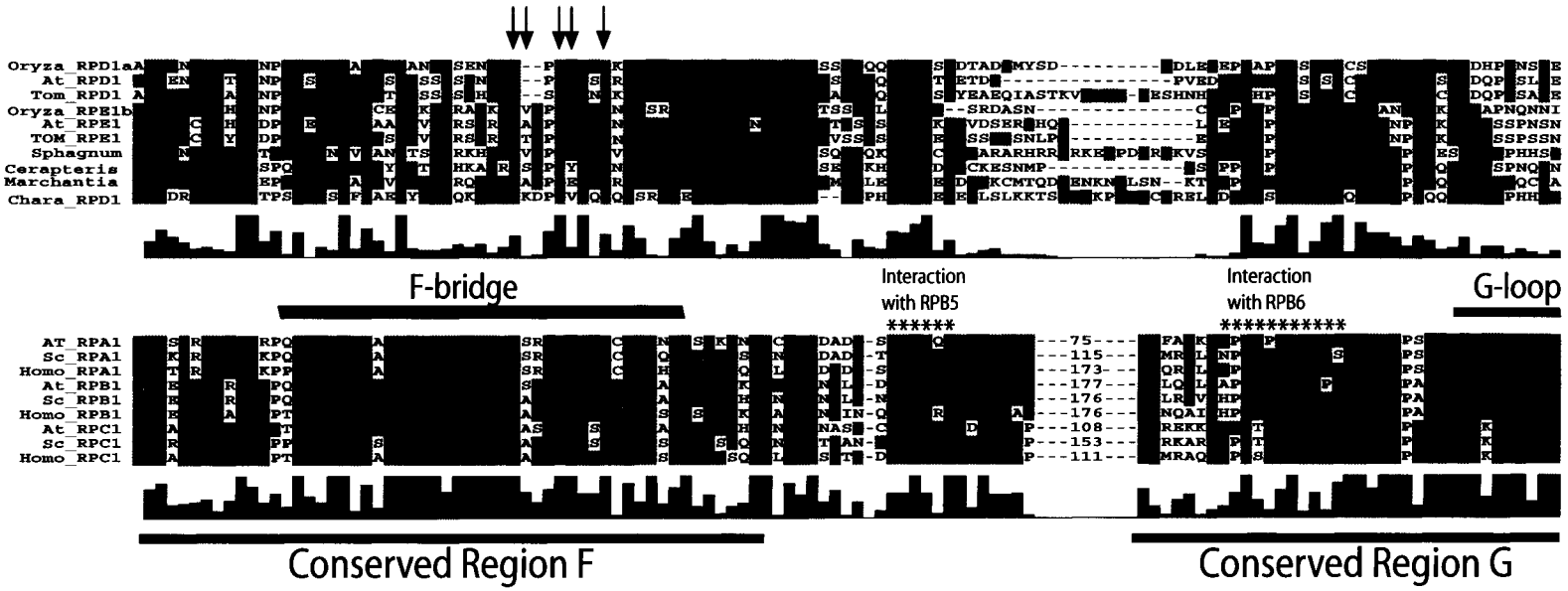


Figure 4.10. The alignment of the largest subunit of RNAPs between regions F and G. There is a complete deletion of sequences between regions F and G in the RPD1/RPE1 sequences. The numbers in the sequences of RNAPs I, II and III show the numbers of residues between the sequences shown. The arrows show the sites that directly contact the template DNA at i+1 and i+2 positions in yeast pol II structure. The stars mark the sites that interact with RPB5 and RPB6, which are also shown as in Fig. 4.6.
The color scheme and histograms are used as described in the legend of Fig 4.6.

Chapter 5. Conclusion and Future Directions

5.1 The evolution of eukaryotic transcription systems

In Chapter 3, I reported the discovery of the Pol IV genes in land plants and charalian green algae and described the genesis of Pol IV from Pol II through a multistep process in which the largest and second-largest subunit genes evolved by independent duplication events (Fig. 3.4 A). Here, I will apply these inferences about the origin of Pol IV to sketch a possible evolutionary history for the origin of three independent polymerase systems in the ancestral eukaryote. I propose that a multistep mechanism similar to that for Pol IV gave rise to the early evolutionary divergence of Pols I, II and III. Initially, gene duplications led to the three largest subunits of Pols I, II and III. Subsequently, duplication of the second-largest subunit gene gave rise to the three second largest subunits of Pols I, II and III (Fig. 3.4B). Prior to duplication of the second largest subunit, the divergent largest subunits of Pols I, II and III shared a common second largest subunit and common small subunits. If this stepwise process did indeed give rise to Pols I, II and III from a single ancestral archaeal-like RNA polymerase in the early eukaryotes, how could the functions of each RNA polymerase then be specified?

Each RNA polymerase in eukaryotes has its unique features and components for promoter recognition, preinitiation complex formation, initiation and promoter clearance, elongation, and termination (Hahn, 2004; White, 2005). Differential recruitment of RNA polymerases occurred at the formation of pre-initiation complexes (PICs) at different

promoters. The Pol II PICs include Pol II (12 subunits), the general transcription factors [GTFs, TATA box binding protein (TBP), TFIIA, -IIB, IIE, -IIF, -IIH, 19 subunits], TBP-associated factors (TAFs, 15 subunits), Mediator (24 subunits), SAGA (~17 subunits), NuA4 (7 subunits), SWI/SNF (~11 subunits), and/or RSC (~11 subunits) (Martinez, 2002; Hahn, 2004). The Pol III PICs include the Pol III (12 subunits), TFIIIC (6 subunits), and TFIIIB [3 subunits, include TBP and BRF (TFIIB homolog)] (Paule and White, 2000). The Pol I PICs are different between mammals and yeast. Besides Pol I (14 subunits), Pol I PICs in mammalian rRNA promoter include UBF, TIF-IB/SL1 (4 subunit with TBP, and three TAFs), and TIF-IA; Pol I PICs in yeast include upstream activating factors (UAF, 6 subunits), TIF-IB (TBP + 3 TAFs), and RRN3 (Grummt, 2003; Ruggero and Pandolfi, 2003). TIF-IA/RRN3 and TBP are the only homologous proteins conserved in mammalian and yeast Pol I PICs. TBP is the only transcription factor that is shared among all three RNA polymerases.

In the formation of Pol II PICs, TFIIB and TBP are responsible for promoter recognition and recruitment of Pol II to form the minimal initiation complex. TBP binds to the TATA box of a Pol II promoter and bends the DNA around the C-terminus of TFIIB. TFIIB recruits Pol II through the N-terminal zinc finger domain (Buratowski and Zhou, 1993; Hahn and Roberts, 2000). TFIIB is also directly involved in promoter opening, open complex formation and start site selection (Bushnell et al., 2004; Chen and Hahn, 2004). In the case of Pol III , Brf, the homolog of TFIIB, is not required for recruitment. Mutations in the Brf zinc ribbon domain inhibit open complex formation, blocking the full opening of DNA strands spanning the transcription start sites (Hahn and

Roberts, 2000; Kassavetis et al., 2001; Kassavetis et al., 2003). However, there is no homolog of TFIIB in Pol I initiation.

In archaea, all genes, transcribed by a single RNA polymerase, have a similar promoter structure (Bell et al., 2001). Transcription in archaebacteria requires only two initiation factors TBP and TFB, for promoter recognition and specific transcription. TBP recognizes the archaeal T/A (TATA) box promoter element and is homologous to TATA-box binding protein (TBP) of pol II transcription (Bell and Jackson, 2000; Geiduschek and Ouhammouch, 2005). TFB is homologous to the TFIIB in pol II transcription. As in pol II transcription initiation, TBP binds to the C-terminal domain of TFB. The zinc finger domain of TFB is required for the recruitment of RNA polymerase and TFB is required for promoter opening and accurate initiation (Bell and Jackson, 2000). Thus, TFB is important for the recruitment and the initiation of RNA polymerase in archaea. Both TBP and TFB play important roles for polymerase recruitment and initiation in archaebacterial and eukaryotic transcription systems, suggesting that the TBP and the TFB were used as recruitment and initiation factors for the single RNAP in the ancestor of eukaryotes, and that this use might have continued after duplication and divergence of the Pol I, II and III largest subunits.

TFIIB interacts with the 'dock' domain in the largest subunit of pol II (Chen and Hahn, 2003). The newly published TFIIB-Pol II crystal structure identifies the region of this interaction to residues 409-419 of *S. cerevisiae* RPB1 (Bushnell et al., 2004). Sequence comparison of the 'dock' domain indicates that the 'dock' domain is conserved

in RPB1 (pol II), RPC1 (pol III) and RPO A1 (archaeal RNAP), in which the TFIIB homologs are used for PICs (Chen and Hahn, 2003). This suggests that similar interactions are used between TFIIB homologs and the 'dock' domains of the largest subunits of pol III and archaeal RNAP. After the initial functional differentiation of Pols I, II and III, it is parsimonious to assume the existence of a specific TFB homolog for Pol I ('TFIB'), Pol II (TFIIB) and Pol III (Brf). Later, as each polymerase system evolved independently and adapted to the new requirements for the gene transcription and regulation, 'TFIB' for Pol I was lost entirely and Brf in Pol III underwent changes that made it no longer important for polymerase recruitment.

Since TFB homologs played similar roles in RNA polymerase recruitment and initiation for the three RNA polymerase during early eukaryote evolution, we assume that divergence of the largest subunits of Pol I, II and III would have been sufficient for functional specification by virtue of having different interactions with TFB homologs (Fig 5.1). Thus, we propose that coevolution of the RPA1/B1/C1 genes and the TFB homologs was correlated with divergence of the three classes of promoters and the evolution of three transcription systems. We can test this evolutionary model of three transcription systems by an *in vitro* RNA polymerase recruitment study. The Pol II PICs can be assembled *in vitro* on a Pol II promoter. We can test whether it is possible to use a TFIIB/Brf hybrid protein to recruit Pol III onto a Pol II promoter, or to use TFIIB to recruit Pol III with a pol II 'dock' domain on RPC1.

5.2 Study of Pol IV function in plants.

The crucial point for understanding the Pol IV function in plants is to know whether Pol IV has RNA polymerase activity. I have tried but failed to demonstrate the presence of a Pol IV RNA polymerase activity in an extract from cauliflower inflorescence. The major challenge was the small amount of pol IV protein in these tissues. In the future, two approaches may be considered for the study of Pol IV activity.

The first way is to choose different tissues to extract Pol IV proteins, especially tissues from monocots such as wheat and rice germ. These, like other plant tissues may prove to have only a small amount of Pol IV. The advantage of using the wheat or rice is that it is easy to distinguish the RNA polymerase activity of Pols I, II and III by α-amanitin. In monocots, Pol I is insensitive to α-amanitin up to 2000ug/ml; Pol II is very sensitive to α-amanitin with half inhibition at 0.05 ug/ml; Pol III has an intermediate sensitivity to α-amanitin with half inhibition at 5ug/ml. This is similar for Pol I and Pol II in cauliflower, but cauliflower Pol III is very resistant to α-amanitin with half inhibition > 500ug/ml and its activity cannot be totally inhibited by α-amanitin (Guifoyle, 1982). The sites for interaction of α-amanitin reside mainly in the conserved region F of the largest subunit (Bushnell et al., 2002) and these sites are no longer conserved in the Pol IV largest subunits. Therefore, if Pol IV has RNA polymerase activity, it is very likely that it will be resistant to inhibition by α-amanitin. Even though Pol IV can be co-purified with Pol III in cauliflower, it is impossible to tell Pol IV activity from Pol III

activity because the pol III activity cannot be completely inhibited by α-amanitin. However, the wheat pol III activity can be completely inhibited by α-amanitin at > 20 ug/ml. So if pol IV is still co-purified with pol III in wheat, it is more likely to identify Pol IV activity from Pol III activity in monocots.

The second approach to study Pol IV activity would be to overexpress the largest and the second largest subunits of Pol IV to get enough protein for analysis. Based on the assumption that the other subunits might be able to be assembled into Pol IV if the largest and the second largest subunits are overexpressed in yeast, I have tried unsuccessfully to overexpress RPD1 and RPD2 in yeast. The reason might be: 1. It is hard to overexpress large proteins in yeast; 2. The codon usage bias may add additional difficulty to overexpress the plant genes in yeast. In the future, baculovirus insect cell expression system could be considered for overexpression of the tagged RPD1 and RPD2 proteins and to test the assembly of Pol IV in insect cells. This experiment is definitely worth trying, and if this experiment of overexpression in insect cells succeeds, it will give information not only about the RNA polymerase activity of Pol IV proteins, but also about RNA polymerase assembly in general. However, since so many factors are uncertain, the success of the experiment is far from certain. An alternative strategy would be to overexpress the largest and second largest subunits of Pol IV in plants. Virus induced overexpression systems and transgene overexpression could be considered for this experimemt. Overexpression of Pol IV subunits in plants would require a more complex experimental design, but it has a greater chance of success than heterologous

expression. For the same reason to distinguish the RNA polymerase activity by α-amanitin, overexpression of Pol IV subunits in Rice may be a good choice.

Pol IV is involved in the RNAi pathway. In order to understand how Pol IV is integrated into the RNAi pathway, we need first to know what kinds of proteins Pol IV associates. Analogously to the C-terminal domain in RPB1, the RPD1 and RPE1 proteins have a C-terminal extension that contains partial repeats. At the end of RPD1 and RPE1 C-terminus, there is a conserved domain, homologous to the DCL protein. DCL is encoded by a nuclear gene, but the protein is imported into chloroplasts. The DCL protein was found to be involved in the maturation of 4.5s rRNA in the chloroplast (Bellaoui et al., 2003). The CTD in RPB1 functions throughout the Pol II transcription cycle. It dynamically associates with proteins involved in transcription, capping, splicing, and polyadenylation (Hirose and Manley, 2000; Howe, 2002; Meinhart et al., 2005). The C-terminal extension in RPD1 and RPE1 may play a similar role in orchestrating the processes involved in the RNAi pathway. The DCL domain in RPD1 and RPE1 may be involved in association of proteins for RNA metabolism. As a future project, one might study the proteins that associate with the DCL domain either through yeast two-hybrid experiments or immunoprecipitation and Mass spectrometry.

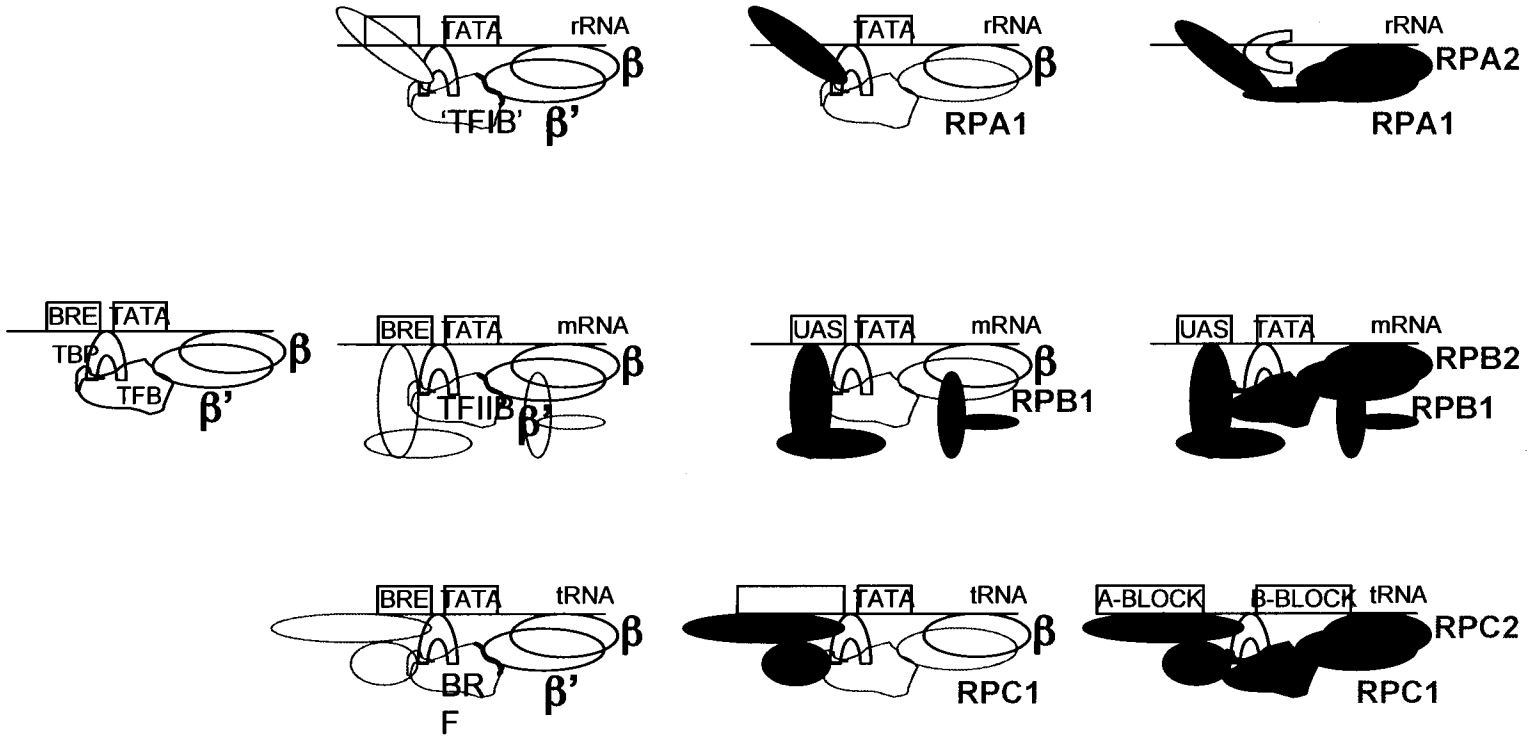


Figure 5.1 A model for the evolution of three eukaryotic transcription systems from a common ancestral transcription system. Initially, there was a triplication of the gene for TFB, which gave rise to "TFIB", TFIIB and BRF (TFIIIB). Subsequently, triplication of RNA polymerase largest subunit gene led to divergence of RPA1, RPB1 and RPC1. Coevolution of the RPA1/B1/C1 and TFB homologs were correlated with the functional divergence of the three RNA polymerases. The "TFIB" in the Pol I system was lost during subsequent evolution. Different colors represent components of different transcription systems.

List of References:

Adman R, Schultz LD, Hall BD (1972) Transcription in yeast: separation and properties of multiple FNA polymerases. Proc Natl Acad Sci U S A **69:** 1702-1706

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796-815

APG (1998) An ordinal classification for the families of flowering plants. Annals of the Missouri Botanical Garden **85:** 531-553

APGII (2003) An update of APG classification for the orders and families of flowering plants. Botanical Journal of the Linnean Society **141:** 399-436

Armache KJ, Mitterweger S, Meinhart A, Cramer P (2005) Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. J Biol Chem **280:** 7131-7134

Becker JD, Boavida LC, Carneiro J, Haury M, Feijo JA (2003) Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol **133:** 713-725

Bell SD, Jackson SP (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends Microbiol **6:** 222-228

Bell SD, Jackson SP (1998) Transcription in Archaea. Cold Spring Harb Symp Quant Biol **63:** 41-51

Bell SD, Jackson SP (2000) The role of transcription factor B in transcription initiation and promoter clearance in the archaeon Sulfolobus acidocaldarius. J Biol Chem **275:** 12934-12940

Bell SD, Magill CP, Jackson SP (2001) Basal and regulated transcription in Archaea. Biochem Soc Trans **29:** 392-395

Bellaoui M, Keddie JS, Gruissem W (2003) DCL is a plant-specific protein required for plastid ribosomal RNA processing and embryo development. Plant Mol Biol **53:** 531-543

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438

Buratowski S, Zhou H (1993) Functional domains of transcription factor TFIIB. Proc Natl Acad Sci U S A **90:** 5633-5637

Bushnell DA, Cramer P, Kornberg RD (2002) Structural basis of transcription: alpha-amanitin-RNA polymerase II cocrystal at 2.8 A resolution. Proc Natl Acad Sci U S A **99:** 1218-1222

Bushnell DA, Westover KD, Davis RE, Kornberg RD (2004) Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science **303:** 983-988

Chan SW, Henderson IR, Jacobsen SE (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat Rev Genet **6:** 351-360

Cheetham GM, Jeruzalmi D, Steitz TA (1999) Structural basis for initiation of transcription from an RNA polymerase-promoter complex. Nature **399:** 80-83

Cheetham GM, Steitz TA (1999) Structure of a transcribing T7 RNA polymerase initiation complex. Science **286:** 2305-2309

Chen HT, Hahn S (2003) Binding of TFIIB to RNA polymerase II: Mapping the binding site for the TFIIB zinc ribbon domain within the preinitiation complex. Mol Cell **12:** 437-447

Chen HT, Hahn S (2004) Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC. Cell **119:** 169-180

Chen X (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science **303:** 2022-2025

Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AM, David PR, Kornberg RD (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. Science **288:** 640-649

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

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

Crawford DJ, Smocovitis VB (2004) The scientific papers of G. Ledyard Stebbins (1929-2000). Koeltz Scientific Books, c2004, Koenigstein, Germany

Dacks JB, Marinets A, Ford Doolittle W, Cavalier-Smith T, Logsdon JM, Jr. (2002) Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. Mol Biol Evol **19:** 830-840

Darwin C (1903) More letters of Charles Darwin, Vol 2. John Murray,, London

Delarue M, Poch O, Tordo N, Moras D, Argos P (1990) An attempt to unify the structure of polymerases. Protein Eng **3:** 461-467

Denton AL, McConaughy BL, Hall BD (1998) Usefulness of RNA polymerase II coding sequences for estimation of green plant phylogeny. Mol Biol Evol **15:** 1082-1085

Dequard-Chablat M, Riva M, Carles C, Sentenac A (1991) RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III). J Biol Chem **266:** 15300-15307

Dietz A, Weisser HJ, Kossel H, Hausmann R (1990) The gene for Klebsiella bacteriophage K11 RNA polymerase: sequence and comparison with the homologous genes of phages T7, T3, and SP6. Mol Gen Genet **221:** 283-286

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG (2001) The highly reduced genome of an enslaved algal nucleus. Nature **410:** 1091-1096

Epshtein V, Mustaev A, Markovtsov V, Bereshchenko O, Nikiforov V, Goldfarb A (2002) Swing-gate model of nucleotide entry into the RNA polymerase active center. Mol Cell **10:** 623-634

Felsenstein J PHYLIP. *In*, Ed 3.6. http://evolution.genetics.washington.edu/phylip.html, Seattle

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:** 496-512

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

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science **296:** 750-752

Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol **3:** 11

Friis EM, Pedersen KR, R. CP (2005) When Earth started blooming: insights from the fossil record. Current Opinion in Plant Biology **8:** 5-12

Geiduschek EP, Ouhammouch M (2005) Archaeal transcription and its regulators. Mol Microbiol **56:** 1397-1407

Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science **292:** 1876-1882

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296:** 92-100

Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol **12:** 1-6

Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood based tests of topologies in phylogenetics. Syst. Biolo. **49:** 652-670

Grant V (1981) Plant Speciation, Ed 2nd. Columbia Univ. Press, New York

Greenleaf AL, Bautz EK (1975) RNA polymerase B from Drosophila melanogaster larvae. Purification and partial characterization. Eur J Biochem **60:** 169-179

Grummt I (2003) Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. Genes Dev **17:** 1691-1702

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. **18:** 2714-2723.

Guilfoyle TJ (1983) DNA-dependent RNA polymerases of plants and lower eukaryotes. Enzymes of Nucleic Acid Synthesis and Modification **II:** 1-42

Guilfoyle TJ, Jendrisak JJ (1978) Plant DNA-dependent RNA polymerases: subunit structures and enzymatic properties of the class II enzymes from quiescent and proliferating tissues. Biochemistry **17:** 1860-1866

Guilfoyle TJ, Lin CY, Chen YM, Key JL (1976) Purification and characterization of RNA polymerase I from a higher plant. Biochim Biophys Acta **418:** 344-357

Hahn S (2004) Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol **11:** 394-403

Hahn S, Roberts S (2000) The zinc ribbon domains of the general transcription factors TFIIB and Brf: conserved functional surfaces but different roles in transcription initiation. Genes Dev **14:** 719-730

Hamilton A, Voinnet O, Chappell L, Baulcombe D (2002) Two classes of short interfering RNA in RNA silencing. Embo J **21:** 4671-4679

Hedtke B, Borner T, Weihe A (1997) Mitochondrial and chloroplast phage-type RNA polymerases in Arabidopsis. Science **277:** 809-811

Hedtke B, Borner T, Weihe A (2000) One RNA polymerase serving two genomes. EMBO Rep **1:** 435-440

Hedtke B, Legen J, Weihe A, Herrmann RG, Borner T (2002) Six active phage-type RNA polymerase genes in Nicotiana tabacum. Plant J **30:** 625-637

Herbeck JT, Wall DP (2005) Converging on a general model of protein evolution. Trends Biotechnol **23:** 485-487

Herr AJ, Jensen MB, Dalmay T, Baulcombe DC (2005) RNA polymerase IV directs silencing of endogenous DNA. Science **308:** 118-120

Hess WR, Borner T (1999) Organellar RNA polymerases of higher plants. Int Rev Cytol **190:** 1-59

Hirose Y, Manley JL (2000) RNA polymerase II and the integration of nuclear events. Genes Dev **14:** 1415-1429

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature **411:** 1046-1049

Honys D, Twell D (2003) Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiol **132:** 640-652

Howe KJ (2002) RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim Biophys Acta **1577:** 308-324

Hu P, Wu S, Sun Y, Yuan CC, Kobayashi R, Myers MP, Hernandez N (2002) Characterization of human RNA polymerase III identifies orthologues for Saccharomyces cerevisiae RNA polymerase III subunits. Mol Cell Biol **22:** 8044-8055

Ishihama A, Kimura M, Mitsuzawa H (1998) Subunits of yeast RNA polymerases: structure and function. Curr Opin Microbiol **1:** 190-196

Jendrisak JJ, Burgess RR (1975) A new method for the large-scale purification of wheat germ DNA-dependent RNA polymerase II. Biochemistry **14:** 4639-4645

Jokerst RS, Weeks JR, Zehring WA, Greenleaf AL (1989) Analysis of the gene encoding the largest subunit of RNA polymerase II in Drosophila. Mol Gen Genet **215:** 266-275

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are moe evolutionarily conserved than are nonessential genes in bacteria. Genome Research **12:** 962-968

Kanno T, Huettel B, Mette MF, Aufsatz W, Jaligot E, Daxinger L, Kreil DP, Matzke M, Matzke AJ (2005) Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nat Genet **37:** 761-765

Karol KG, McCourt RM, Cimino MT, Delwiche CF (2001) The closest living relatives of land plants. Science **294:** 2351-2353

Kassavetis GA, Han S, Naji S, Geiduschek EP (2003) The role of transcription initiation factor IIIB subunits in promoter opening probed by photochemical cross-linking. J Biol Chem **278:** 17912-17917

Kassavetis GA, Letts GA, Geiduschek EP (2001) The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. Embo J **20:** 2823-2834

Kato H, Goto DB, Martienssen RA, Urano T, Furukawa K, Murakami Y (2005) RNA polymerase II is required for RNAi-dependent heterochromatin assembly. Science **309:** 467-469

Kettenberger H, Armache KJ, Cramer P (2004) Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. Mol Cell **16:** 955-965

Kimura M (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc Natl Acad Sci U S A **88:** 5969-5973

Kimura M, Ishiguro A, Ishihama A (1997) RNA polymerase II subunits 2, 3, and 11 form a core subassembly with DNA binding activity. J Biol Chem **272:** 25851-25855

Kimura M, Ohta T (1974) On some principles governing molecular Evolution. Proc Natl Acad Sci U S A **71:** 2848-2852

Korzheva N, Mustaev A, Kozlov M, Malhotra A, Nikiforov V, Goldfarb A, Darst SA (2000) A structural model of transcription elongation. Science **289:** 619-625

Kramer EM, Dorit RL, Irish VF (1998) Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. Genetics **149:** 765-783

Kramer EM, Jaramillo MA, Di Stilio VS (2004) Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. Genetics **166:** 1011-1023

Langer D, Hain J, Thuriaux P, Zillig W (1995) Transcription in archaea: similarity to that in eucarya. Proc Natl Acad Sci U S A **92:** 5768-5772

Larkin RM, Guilfoyle TJ (1996) A 14-kDa Arabidopsis thaliana RNA polymerase III subunit contains two alpha-motifs flanked by a highly charged C terminus. Gene **172:** 211-215

Larkin RM, Hagen G, Guilfoyle TJ (1999) Arabidopsis thaliana RNA polymerase II subunits related to yeast and human RPB5. Gene **231:** 41-47

Lippman Z, Martienssen R (2004) The role of RNA interference in heterochromatic silencing. Nature **431:** 364-370

Litt A, Irish VF (2003) Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. Genetics **165:** 821-833

Liu YJ, Hall BD (2004) Body plan evolution of ascomycetes, as inferred from an RNA polymerase II phylogeny. Proc Natl Acad Sci U S A **101:** 4507-4512

Liu YJ, Whelen S, Hall BD (1999) Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit. Mol Biol Evol **16:** 1799-1808

Ma X-F, Gustafson JP (2005) Genome evolution of allopolyploids: a process of cytological and genetical diploidization. Cytogenet. Genome Res **109:** 236-249

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A **102:** 5454-5459

Martinez E (2002) Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol **50:** 925-947

Masterson J (1994) STOMATAL SIZE IN FOSSIL PLANTS - EVIDENCE FOR POLYPLOIDY IN MAJORITY OF ANGIOSPERMS. Science **264:** 421-424

Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264:** 421-423

McCourt RM, Delwiche CF, Karol KG (2004) Charophyte algae and land plant origins. Trends in Ecology and Evolution **19:** 661-666

McKune K, Moore PA, Hull MW, Woychik NA (1995) Six human RNA polymerase subunits functionally substitute for their yeast counterparts. Mol Cell Biol **15:** 6895-6900

Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P (2005) A structural perspective of CTD function. Genes Dev **19:** 1401-1415

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A **102:** 10930-10935

Miyao T, Woychik NA (1998) RNA polymerase subunit RPB5 plays a role in transcriptional activation. Proc Natl Acad Sci U S A **95:** 15281-15286

Mooney RA, Landick R (1999) RNA polymerase unveiled. Cell **98:** 687-690

Mooney RA, Landick R (1999) RNA polymerase unveiled. Cell **98:** 687-690

Nickerson J, Drouin G (2004) The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. Mol Phylogenet Evol **31:** 403-415

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

Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet **34:** 401-437

Oxelman B, Yoshikawa N, McConaughy BL, Luo J, Denton AL, Hall BD (2004) RPB2 gene phylogeny in flowering plants, with particular emphasis on asterids. Mol Phylogenet Evol. **32:** 462-479

Pal-Bhadra M, Leibovitch BA, Gandhi SG, Rao M, Bhadra U, Birchler JA, Elgin SC (2004) Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science **303:** 669-672

Paule MR, White RJ (2000) Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res. **28:** 1283-1298

Pong SS, Loomis WF, Jr. (1973) Multiple nuclear ribonucleic acid polymerases during development of Dictyostelium discoideum. J Biol Chem **248:** 3933-3939

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics **14:** 817-818

Pryer KM, Schneider H, Smith AR, Cranfill R, Wolf PG, Hunt JS, Sipes SD (2001) Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature **409:** 618-622

Puhler G, Leffers H, Gropp F, Palm P, Klenk HP, Lottspeich F, Garrett RA, Zillig W (1989) Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. Proc Natl Acad Sci U S A **86:** 4569-4573

Puhler G, Lottspeich F, Zillig W (1989) Organization and nucleotide sequence of the genes encoding the large subunits A, B and C of the DNA-dependent RNA polymerase of the archaebacterium Sulfolobus acidocaldarius. Nucleic Acids Res **17:** 4517-4534

Rambaut A (1996) Se-Al: Sequence Alignment Editor. *In*, Ed 2.0. http://evolve.zoo.ox.ac.uk/software.html?id=seal

Ramsey J, Schemske DW (1998) Pathways, mechanism and rates of polyploid formation in flowering plants. Annul Rev. Ecol. Syst. **29:** 467-501

Richter U, Kiessling J, Hedtke B, Decker E, Reski R, Borner T, Weihe A (2002) Two RpoT genes of Physcomitrella patens encode phage-type RNA polymerases with dual targeting to mitochondria and plastids. Gene **290:** 95-105

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

Ronquist F, Huelsenbeck J (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574

Ruggero D, Pandolfi PP (2003) Does the ribosome translate cancer? Nat Rev Cancer **3:** 179-192

Sakurai H, Miyao T, Ishihama A (1996) Subunit composition of RNA polymerase II from the fission yeast Schizosaccharomyces pombe. Gene **180:** 63-67

Sentenac A (1985) Eukaryotic RNA polymerases. CRC Crit Rev Biochem **18:** 31-90

Serino G, Maliga P (1998) RNA polymerase subunits encoded by the plastid rpo genes are not shared with the nucleus-encoded plastid enzyme. Plant Physiol **117:** 1165-1170

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol **16:** 1114-1116

Shpakovski GV, Acker J, Wintzerith M, Lacroix JF, Thuriaux P, Vigneron M (1995) Four subunits that are shared by the three classes of RNA polymerase are functionally interchangeable between Homo sapiens and Saccharomyces cerevisiae. Mol Cell Biol **15:** 4702-4710

Shultz JW, Regier JC (2000) Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. Proc Biol Sci **267:** 1011-1019

Sklar VE, Jaehning JA, Gage LP, Roeder RG (1976) Purification and subunit structure of deoxyribonucleic acid-dependent ribonucleic acid polymerase III from the posterior silk gland of Bombyx mori. J Biol Chem **251:** 3794-3800

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci U S A **97:** 7051-7057

Sosunova E, Sosunov V, Kozlov M, Nikiforov V, Goldfarb A, Mustaev A (2003) Donation of catalytic residues to RNA polymerase active center by transcription factor Gre. Proc Natl Acad Sci U S A **100:** 15469-15474

Sousa R, Chung Y, Rose J, Wang B (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. Nature **364:** 593-599

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Stebbins GL (1971) Chromosomal evolution in Higher Plants. Addison-Wesley, New York

Stefanovic S, Olmstead RG (2004) Testing the phylogenetic position of a parasitic plant (Cuscuta, Convolvulaceae, asteridae): Bayesian inference and the parametric bootstrap on data drawn from three genomes. Syst Biol **53:** 384-399

Steitz TA, Smerdon S, Jager J, Wang J, Kohlstaedt LA, Friedman JM, Beese LS, Rice PA (1993) Two DNA polymerases: HIV reverse transcriptase and the Klenow fragment of Escherichia coli DNA polymerase I. Cold Spring Harb Symp Quant Biol **58:** 495-504

Steitz TA, Smerdon SJ, Jager J, Joyce CM (1994) A unified polymerase mechanism for nonhomologous DNA and RNA polymerases. Science **266:** 2022-2025

Stiller JW, Duffield EC, Hall BD (1998) Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. Proc Natl Acad Sci U S A **95:** 11769-11774

Stiller JW, Hall BD (1997) The origin of red algae: implications for plastid evolution. Proc Natl Acad Sci U S A **94:** 4520-4525

Sweetser D, Nonet M, Young RA (1987) Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. Proc Natl Acad Sci U S A **84:** 1192-1196

Swofford D (2002) PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta. *In*,

Tahirov TH, Temiakov D, Anikin M, Patlan V, McAllister WT, Vassylyev DG, Yokoyama S (2002) Structure of a T7 RNA polymerase elongation complex at 2.9 A resolution. Nature **420:** 43-50

Temiakov D, Patlan V, Anikin M, McAllister WT, Yokoyama S, Vassylyev DG (2004) Structural basis for substrate selection by t7 RNA polymerase. Cell **116:** 381-391

Thompson JD, Gilbson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. Nucleic Acid Res. 25:4876-82

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, Nelson K, Quackenbush J, Zhou L, Kirkness EF, Peterson S, Loftus B, Richardson D, Dodson R, Khalak HG, Glodek A, McKenney K, Fitzegerald LM, Lee N, Adams MD, Hickey EK, Berg DE, Gocayne JD, Utterback TR, Peterson JD, Kelley JM, Cotton MD, Weidman JM, Fujii C, Bowman C, Watthey L, Wallin E, Hayes WS, Borodovsky M, Karp PD, Smith HO, Fraser CM, Venter JC (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature **388:** 539-547

Ulmasov T, Larkin RM, Guilfoyle TJ (1995) Arabidopsis expresses two genes that encode polypeptides similar to the yeast RNA polymerase I and III AC40 subunit. Gene **167:** 203-207

Ulmasov T, Larkin RM, Guilfoyle TJ (1996) Association between 36- and 13.6-kDa alpha-like subunits of Arabidopsis thaliana RNA polymerase II. J Biol Chem **271:** 5085-5094

Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, Borukhov S, Yokoyama S (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. Nature **417:** 712-719

Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA (2002) Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science **297:** 1833-1837

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102:** 5483-5488

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol **42:** 225-249

Westover KD, Bushnell DA, Kornberg RD (2004) Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. Cell **119:** 481-489

Westover KD, Bushnell DA, Kornberg RD (2004) Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. Science **303:** 1014-1016

White MJD (1978) Modes of speciation. W. H. Freeman, New York, New York, USA

White RJ (2005) RNA polymerase I and III, growth control and cancer. Nat Rev Mol Cell Biol **6:** 69-78

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann. Rev. Biochem **46**

Yao MC, Fuller P, Xi X (2003) Programmed DNA deletion as an RNA-guided system of genome defense. Science **300:** 1581-1584

Yin YW, Steitz TA (2004) The structural mechanism of translocation and helicase activity in T7 RNA polymerase. Cell **116:** 393-404

Young HA, Whiteley HR (1975) Deoxyribonucleic acid-dependent ribonucleic acid polymerases in the dimorphic fungus Mucor rouxii. J Biol Chem **250:** 479-487

Young RA (1991) RNA polymerase II. Annu Rev Biochem **60:** 689-715

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science **296:** 79-92

Zhang G, Campbell EA, Minakhin L, Richter C, Severinov K, Darst SA (1999) Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell **98:** 811-824

Zilberman D, Cao X, Jacobsen SE (2003) ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science **299:** 716-719

Zillig W, Klenk HP, Palm P, Puhler G, Gropp F, Garrett RA, Leffers H (1989) The phylogenetic relations of DNA-dependent RNA polymerases of archaebacteria, eukaryotes, and eubacteria. Can J Microbiol **35:** 73-80

Vita

1998-2006	**University of Washington, WA** Ph.D., Department of Biology
1995-1998	**Shanghai Institute of Plant Physiology, The Chinese Academy of Sciences, China** M.Sc., Plant Physiology
1991-1995	**Sichuan University, China** B.Sc., Department of Biology.