Suggested running title (not more than 60 characters, including spaces):
Expansion of poplar MYB family

Corresponding Author:
Malcolm M. Campbell
Department of Cell & Systems Biology
University of Toronto
25 Willcocks St.
Toronto, Ontario M5S 3B2
Canada
voice: (+1) 416 946 0817
fax: (+1) 416 978 5878
email: malcolm.campbell@utoronto.ca

Research Category:
Genetics, Genomics and Molecular Evolution
Full title of the paper:

Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors

Full names of all the authors:

Olivia Wilkins\textsuperscript{1,2}, Hardeep Nahal\textsuperscript{1,2}, Justin Foong\textsuperscript{1,2}, Nicholas J. Provar\textsuperscript{1,2} & Malcolm M. Campbell\textsuperscript{1,2}

Names and addresses of the institutions:

\textsuperscript{1}Department of Cell & Systems Biology, and \textsuperscript{2}Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks St., Toronto, ON M5S 3B2, Canada
Funding:
Research infrastructure and technical support was generously provided by the Centre for Analysis of Genome Evolution and Function at University of Toronto. OW was generously supported by a Natural Science and Engineering Research Council of Canada (NSERC) Canadian Graduate Scholarship (CGSD). This work was generously supported by funding from NSERC and the Canada Foundation for Innovation (CFI) to NJP, and by funding from the University of Toronto, CFI, and NSERC to MMC.

Present Addresses:
Same as above

Corresponding Author:
Malcolm M. Campbell
Malcolm.campbell@utoronto.ca
The R2R3-MYB proteins comprise one of the largest families of transcription factors in plants. R2R3-MYB family members regulate plant-specific processes, such as the elaboration of specialised cell types, including xylem, guard cells, trichomes and root hairs, and the biosynthesis of specialised branches of metabolism, including phenylpropanoid biosynthesis. As such, R2R3-MYB family members are hypothesised to contribute to the emergence of evolutionary innovations that have arisen in specific plant lineages. As a first step in determining the role played by R2R3-MYB family members in the emergence of lineage-specific innovations in the genus *Populus*, the entire *Populus trichocarpa* R2R3-MYB family was characterised. The *Populus* R2R3-MYB complement is much larger than that found in other angiosperms with fully sequenced genomes. Phylogenetic analyses, together with chromosome placement, showed that the expansion of the *Populus* R2R3-MYB family was not only attributable to whole genome duplication, but also involved selective expansion of specific R2R3-MYB clades. Expansion of the *Populus* R2R3-MYB family prominently involved members with expression patterns that suggested a role in specific components of *Populus* life history, including wood formation and reproductive development. An expandable compendium of microarray-based expression data (PopGenExpress) and associated web-based tools were developed to better enable within- and between-species comparisons of *Populus* R2R3-MYB gene expression. This resource, which includes intuitive graphic visualisation of gene expression data across multiple tissues, organs and treatments, is freely available to, and expandable by, scientists wishing to better understand the genome biology of *Populus*, an ecologically dominant and economically important forest tree genus.
INTRODUCTION

Plant growth and development are regulated by the coordinated expression of tens of thousands of genes. This is achieved through the actions of transcription factors - proteins that show sequence specific DNA binding that activate or repress transcription in response to endogenous and exogenous stimuli (Riechmann et al., 2000). To accommodate the demands of a sessile life form, plants have evolved large families of transcription factors (Riechmann et al., 2000). In the Arabidopsis thaliana genome, approximately 6% of the estimated total number of genes are transcription factors of which an estimated 9% encode members of the MYB family (Riechmann et al., 2000).

The MYB gene family is at least one billion years old and is represented in genomes in all major eukaryotic lineages (Lipsick, 1996; Rosinski and Atchley, 1998; Kranz et al., 2000). The MYB proteins are named for the v-MYB oncogene, a determinant of avian myeloblastosis that was first discovered in chickens (Lipsick, 1996). MYB proteins are identified on the basis of a highly conserved amino-terminal DNA-binding domain (DBD) (Lipsick, 1996). The MYB DBD consists of imperfect repeats of an approximately 50 amino acid region, with each of the repeats giving rise to a helix-helix-loop-helix secondary structure that binds to the major groove of the target DNA sequences (Lipsick, 1996; Stracke et al., 2001). The MYB DBD contains several highly conserved tryptophan residues, which play a role in sequence specific binding of DNA (Ogata et al., 1995). In contrast, the region C-terminal to the DNA binding domain is the activation domain, and varies considerably between MYB proteins (Kranz et al., 1998; Jin and Martin, 1999; Stracke et al., 2001; Jiang et al., 2004). This speaks to the wide range of regulatory roles played by members of this family (Kranz et al., 1998; Jin and Martin, 1999; Meissner et al., 1999).

All known animal MYB DBDs consist of three tandem repeats of the MYB motif (designated R1, R2, and R3) and are referred to as 3R-MYB proteins (Rosinski and Atchley, 1998). Plant MYBs may contain three MYB repeats, but more commonly they have only two (R2R3) or one (R1) MYB DBD (Kranz et al., 1998; Kranz et al., 2000; Chen et al., 2006). Furthermore, the MYB family is greatly expanded in plant genomes relative to animal genomes (Stracke et al., 2001). In all known vertebrate genomes, there are three different 3R-MYB proteins – c-MYB, A-MYB and B-MYB (Lipsick, 1996; Rosinski and Atchley, 1998). In contrast, in the A. thaliana genome, there are 5 3R-MYB genes, and 126 R2R3-MYB genes (Stracke et al., 2001; Durberry et al., 2005), while in rice there are 84 R2R3-MYB genes (Jiang et al., 2004).
This expansion of the R2R3-MYB family in plants supports the observation that many R2R3-MYB transcription factors play roles in plant specific processes. For example, the elaboration of specific epidermal cell types, including trichomes (Oppenheimer et al., 1991; Higginson et al., 2003), root hairs (Wada et al., 1997; Tominaga et al., 2007), and petal epidermis (Perez-Rodriguez et al., 2005; Baumann et al., 2007) is contingent on the activity of R2R3-MYB transcription factors. R2R3-MYB family members also play central roles in the control of plant metabolism, notably the regulation of synthesis of specific classes of phenylpropanoids, such as flavonoids (Mehrtens et al., 2005; Stracke et al., 2007), anthocyanin (Teng et al., 2005; Deluc et al., 2006) and lignins (Patzlaff et al., 2003; Goicoechea et al., 2005; Bedon et al., 2007). R2R3-MYB family members have demonstrated roles in regulating plant response to phytohormonal cues including abscisic acid (Abe et al., 1997; Abe et al., 2003) and gibberilllin (Gocal et al., 2001) and to environmental signals such as light (Jin et al., 2000; Hartmann et al., 2005) water availability (Urao et al., 1996; Jung et al., 2008).

Given the multiplicity of plant-specific processes controlled by R2R3-MYB transcription factors, it has been postulated that the elaboration of the R2R3-MYB family may account for some of the evolutionary innovations that contribute to plant diversity (Riechmann et al., 2000). In A. thaliana, members of the R2R3-MYB family of transcription factors have been identified and their targets and expression patterns are being characterised (Kranz et al., 1998; Meissner et al., 1999; Stracke et al., 2001). However, A. thaliana is an herbaceous, annual plant, which is not representative of a large number of important plant species, particularly long-lived woody perennial plants, like forest trees. Trees are distinct from herbaceous species in many ways: they have a self-supporting structure, the secondary growth or wood, and they have a much longer lifespan (Bradshaw et al., 2000; Brunner et al., 2004). The regulatory networks and molecular mechanisms that underlie these unique properties cannot be investigated through the examination of non-tree species, thus demonstrating the importance and opportunity presented by the recent publication of the complete genome sequences for two wood-forming species – Populus trichocarpa (Tuskan et al., 2006), and Vitis vinifera (Velasco et al., 2007).

Molecular evidence suggests that P. trichocarpa and A. thaliana shared their last common ancestor some 100 to 120 million years ago (Tuskan et al., 2006). Since then, A. thaliana and P. trichocarpa have evolved different life histories, including herbaceous versus arboreal development, annual versus perennial habit, and selfing versus outcrossing mating strategies. In both plants and animals, the evolution of such diversity has been hypothesised to be the consequence of changes in
the expression pattern of the genes encoding transcription factors, and/or changes in the function of
the transcription factors themselves (Doebley and Lukens, 1998; Weatherbee et al., 1998). The
elaboration of the R2R3-MYB transcription factor family is postulated to account for some of the
evolutionary innovations that contribute to plant diversity (Jin and Martin, 1999). Here we explore
the contribution of R2R3-MYB proteins in the diversification of plant form and function, through an
analysis of the entire R2R3-MYB family encoded in the *Populus* genome.

RESULTS AND DISCUSSION

The *Populus* R2R3-MYB family is larger than that of other sequenced dicotyledonous angiosperms

The *Populus* R2R3-MYB family was compared to the corresponding families from the woody perennial *Vitis vinifera*, which is a sister taxon to *Populus* in Eurosids I, and the more distantly related *A. thaliana*, which is a member of Eurosids II. The predicted R2 and R3 MYB repeats of the MYB DBD are highly conserved across plant lineages (Jin and Martin, 1999). Using the conserved amino-terminal DBD as the defining feature, 192 genes in the *P. trichocarpa* genome were annotated as encoding R2R3-MYB proteins, and five genes were annotated as encoding 3R-MYB proteins (Table I, Supplemental Table S1). Like their counterparts in other plant species, the R2 and R3 MYB repeats of the *P. trichocarpa* R2R3-MYB family contain characteristic amino acids, including a series of evenly distributed and highly conserved tryptophan residues that play a role in sequence-specific binding of DNA (Ogata et al., 1995; Stracke et al., 2001) (Fig. 1). In keeping with this, the phylogeny constructed with all predicted *Populus 3R-MYB* and R2R3-MYB proteins, plus all *A. thaliana* and *V. vinifera* family members (Fig. 2), is highly similar to a previously published phylogeny which included all known *A. thaliana* and *Oryza sativa* R2R3-MYB proteins (Jiang et al., 2004). Most of the subgroups described in this previously published phylogeny, defined on the basis of their *A. thaliana* complement, are maintained in the current phylogeny, with most subgroups merely expanding to include both *Populus* and *Vitis* R2R3-MYB members (Supplemental Table S2).

The *P. trichocarpa* genome encodes many more R2R3-MYB family members (192) than either *A. thaliana* (126) (Stracke et al., 2001; Durbary et al., 2005) or *Vitis vinifera* (123). This expansion appears to be the result of multiple gene duplication processes including a whole-genome duplication event in the *Populus* lineage as well as multiple segmental and tandem duplication events (Tuskan et al., 2006). To examine the relative contribution of each of these factors in the
expansion of the R2R3-MYB family, R2R3-MYB genes were electronically mapped to loci across all 19 linkage groups (LG). In total, 172 *Populus* R2R3-MYB genes were mapped, it was not possible to map the remaining 21 R2R3-MYB genes as they have not yet been assigned to any linkage group. There are as many as 21 (LG_II) and as few as a single (LG_XVI) R2R3-MYB genes per linkage group. On average, there is one R2R3-MYB gene every 2.4Mb.

**Expansion of the *Populus* R2R3-MYB family has occurred through multiple mechanisms, and resulted in lineage-specific differences in R2R3-MYB clades**

Evidence of the salicoid-specific whole genome duplication event in the *Populus* lineage (ca. 65 MYA) is present throughout the *P. trichocarpa* genome (Tuskan et al., 2006). Consistent with this, many predicted MYB-encoding genes have paralogous counterparts in syntenic regions of related linkage groups. For example, large sections of linkage groups II and V are thought to be the product of the salicoid-specific genome duplication (Tuskan et al., 2006). In keeping with this, three predicted MYB genes on linkage group II (LG_II) (*Ptr*MYB109, *Ptr*MYB189, and *Ptr*MYB220) show conserved organization with three highly similar MYB genes on LG_V (*Ptr*MYB030, *Ptr*MYB158, and *Ptr*MYB190). Many similar examples exist throughout the *Populus* MYB gene family (Fig. 3); however, whole genome duplication can only explain some of the R2R3-MYB gene family expansion observed.

Tandem gene duplication has also played an important role in the elaboration of the R2R3-MYB gene family. More than 35% (68/192) of the R2R3-MYB-encoding genes in the *P. trichocarpa* genome are present as tandem repeats, where the gene duplications were directly adjacent to each other on a given linkage group, with no intervening annotated gene. Tandem repeats most commonly include duplicate or triplicate R2R3-MYB-encoding genes, but there is one instance of four tandem repeats (LG_XI) and one of six repeats (LG_II).

While the number of R2R3-MYB-encoding genes has expanded in *Populus*, the number of 3R-MYB genes has not. That is, *A. thaliana* (Stracke et al., 2001), *V. vinifera* and *P. trichocarpa* genomes all encode only five 3R-MYB proteins, despite the fact that they encode a widely varying number of R2R3-MYB proteins – 125, 123 and 192 respectively (Supplemental Tables S1, S3, S4). The 3R-MYB gene typifies the vertebrate MYB family (Lipsick, 1996) and is present in all major plant lineages (Kranz et al., 2000; Ito, 2005; Haga et al., 2007) suggesting that 3R-MYB genes represent an ancient and evolutionarily conserved gene family. In both plants and vertebrates, 3R-MYB...
proteins regulate progress through cell cycle transition (Ito et al., 2001; Okada et al., 2002; Ito, 2005; Haga et al., 2007). Though there appears to be partially redundant function within this family of genes in *A. thaliana*, loss-of-function mutations in 3R-MYB genes have been demonstrated to lead to incomplete cytokinesis (Haga et al., 2007). This suggests that 3R-MYB proteins fulfill a number of core cellular functions that are conserved across evolution, while the larger, more diverse family of R2R3-MYB proteins in plants suggests that they may play a role in generating phenotypic diversity within this kingdom.

Phylogenetic analysis of the predicted R2R3-MYB protein sequences revealed that there was not equal representation of *Populus*, *Vitis* and *A. thaliana* R2R3-MYB proteins within given clades (Fig. 2, Supplemental Fig. S1). For example, phylogeny subgroup C25, which includes only one *A. thaliana* MYB proteins and two *Vitis* MYB proteins, includes seven *Populus* R2R3-MYB proteins. By contrast, the C13 subgroup includes six *A. thaliana* and two *Vitis* MYB proteins, but not one predicted *Populus* MYB protein. In other cases, such as subgroup C22, there are fewer *Populus* MYB proteins than expected, given the whole genome duplication event (five *A. thaliana*, three *P. trichocarpa*, three *V. vinifera*). These findings lend support to the model of gene loss in the *Populus* lineage or gene duplication in the *A. thaliana* lineage following divergence from the last common ancestor.

Remarkably, several clades do not include any *A. thaliana* R2R3-MYB proteins, but only members from *Populus* and *Vitis* (Fig. 2). This suggests that the genes in these clades may have specialised roles that were either lost in *A. thaliana* or acquired in the *Populus* and *Vitis* lineages after divergence from the last common ancestor with *A. thaliana*. It remains to be determined whether the absence of *A. thaliana* genes in these clades extends to other members of Eurosids II, or whether it is something particular to the *A. thaliana* genome.

**PopGenExpress**, an Affymetrix microarray-based compendium of *Populus* transcript abundance, provides insights into R2R3-MYB gene expression and diversification

Affymetrix Poplar Genome Arrays were used to assess the transcript abundance of 180 R2R3-MYB-encoding genes (Supplemental Table S1). There were no probe sets on the array corresponding to the remaining R2R3-MYB-encoding transcripts. Transcript abundance for the R2R3-MYB-encoding genes was assessed in biological triplicate RNA samples extracted from seedlings grown under different light regimes, young leaves, mature leaves, roots, xylem, female catkins, and male catkins.
The compendium of data derived from these experiments is referred to as the *Populus* gene expression (PopGenExpress) dataset.

As is commonly the case for genes encoding transcription factors, many of the *P. trichocarpa* R2R3-MYB-encoding genes had low transcript abundance levels as determined by the Affymetrix microarray analysis. Nevertheless, distinct transcript abundance patterns were readily identifiable in the PopGenExpress dataset for all 180 of the R2R3-MYB probe sets on the microarray. Groups of MYB-encoding genes showed preferential accumulation of transcripts in a given organ or tissue, or under a specific condition (Fig. 4). In fact, the majority (75%) of *Populus* R2R3-MYB-encoding genes exhibited transcript abundance profiles that had marked peaks in transcript abundance in only one distinct condition in the current PopGenExpress dataset. This suggests that R2R3-MYB proteins function as regulators of processes that are limited to discrete cells, organs or conditions.

In keeping with their roles as regulators in plant specific processes, 23/180 *Populus* R2R3-MYB-encoding genes showed the highest level of transcript abundance in differentiating xylem, a tissue that gives rise to the woody stem characteristic of trees like *Populus*. A further 29 R2R3-MYB genes (16%) had the greatest accumulation of transcripts in roots. Remarkably, none of these genes encoded a protein with high sequence similarity to WEREWOLF, which is involved in the determination of epidermal cell fate in *A. thaliana* roots (Tominaga et al., 2007). It may be that one of the other R2R3-MYB genes with high transcript abundance in *Populus* roots plays this role.

Strikingly, seven of the *Populus* MYB family members with most abundant expression in roots encode proteins in a clade with members of the Blind subfamily of R2R3-MYB proteins (C1). Blind was originally identified on the basis of its role in affecting plant aerial architecture by controlling axillary branch formation in tomato (Schmitz et al., 2002). Blind-related proteins in *A. thaliana* have demonstrated roles in the regulation of axillary meristems (*AtMYB37, AtMYB38, AtMYB84*) (Keller et al., 2006; Muller et al., 2006) and in root development (*AtMYB68*) (Feng et al., 2004). It may be that the Blind-related proteins that are expressed in *Populus* roots play a similar role in shaping root architecture. Similarly, it is possible that some of the Blind-related genes in *P. trichocarpa* may exhibit transcript accumulation in axillary buds and meristems. The continued expansion of the PopGenExpress data set will allow for testing of such hypotheses.

Approximately 28% (50/180) of the *Populus* R2R3-MYB-encoding genes had highest transcript abundance in catkins. Twenty-nine of these fifty genes (58%) showed the highest transcript
accumulation in male catkins, seven (14%) had the highest accumulation in female catkins and the remaining fourteen (28%) had approximately equal transcript accumulation in male and female catkins. The genes showing highest transcript accumulation in both male and female catkins represent the largest group of R2R3-MYB transcripts that accumulate in more than one condition in this data set.

Twelve of the R2R2-MYB-encoding genes with highest transcript accumulation in flowers fell into four clades (C11, C19, C38, C39) that contained A. thaliana R2R3-MYB family members with known roles in plant reproductive biology. These included A. thaliana R2R3-MYB family members implicated in anther development (AtMYB21 (Shin et al., 2002), AtMYB24 (Yang et al., 2007), AtMYB33 (Gocal et al., 2001), AtMYB65 (Gocal et al., 2001), AtMYB103 (Zhang et al., 2007)), pollen tube guidance and synergid formation (AtMYB98) (Kasahara et al., 2005), and sperm cell formation (AtMYB125) (Durbarry et al., 2005). The Populus counterparts of these MYB family members, which share both high sequence similarity and expression profiles, are good candidates for future studies aimed at developing a better mechanistic understanding of floral development in Populus. Characterisation of these family members would be an ideal starting point in the dissection of sex determination in the dioecious Populus lineage.

In addition to groups of genes that had similar transcript abundance profiles but were relatively phylogenetically distinct, several phylogenetic clades were characterised by having members that largely shared the same transcript abundance profile. Similar transcript abundance patterns were observed even between A. thaliana and Populus members of the clade. Prominent amongst these clades were those that included R2R3-MYB-encoding genes related to those previously implicated in the regulation of phenylpropanoid metabolism.

Populus R2R3-MYB proteins related to those implicated in regulation of phenylpropanoid metabolism tend to share transcript abundance profiles

Many R2R3-MYB proteins modulate expression of genes encoding enzymes involved in various facets of phenylpropanoid metabolism. Phenylpropanoid metabolism generates a vast array of compounds that are important for a diversity of plant functions, including resistance to herbivore and pathogen attacks (Peters and Constabel, 2002; Miranda et al., 2007), and for the construction of structural components of the plant body (Boerjan et al., 2003; Deluc et al., 2006). Phenylpropanoid metabolism is particularly important in tree species, which make use of phenylpropanoid-derived
metabolites to elaborate a diversity of soluble defence compounds as well as insoluble tannins that function as a durable resistance against herbivores and pathogens (Peters and Constabel, 2002). Moreover, trees like *Populus* have a woody stem composed of up to 30% phenylpropanoid-derived lignins, which provide structural integrity, water transport capacity, and defence against degradation for this key portion of the tree body. In keeping with the importance of maintaining a diverse and robust phenylpropanoid metabolism in trees, genes encoding enzymes of phenylpropanoid metabolism are over-represented in the *Populus* genome relative to the genome *A. thaliana* (Tuskan et al., 2006). On the basis of the analysis of *Populus* genes predicted to encode R2R3-MYB proteins, the trend in diversifying genes encoding enzymes in the phenylpropanoid pathway also seems to extend to genes encoding regulators of this pathway.

Many of the R2R3-MYB proteins implicated in the control of phenylpropanoid metabolism group into specific clades on the basis of the facet of phenylpropanoid metabolism they regulate. For example, clade number ten (C10) includes R2R3-MYB family members implicated in regulation of genes encoding lignin biosynthetic enzymes (Fig. 5a). This clade includes *Pinus taeda* MYB4 (*PtrMYB4*), which is expressed in cells undergoing lignification, and alters the accumulation of transcripts corresponding to genes encoding lignin biosynthetic enzymes (Patzlaff et al., 2003). Other members of this clade have been shown to function in an almost identical fashion with expression in differentiating xylem cells and capacity to alter lignin biosynthesis. This includes *AfMYB46* (Zhong et al., 2007), *Eucalyptus gunnii* MYB2 (*EgMYB2*) (Goicoechea et al., 2005), and *Picea glauca* MYB4 (*PgMYB4*) (Bedon et al., 2007).

In the present phylogeny, two *V. vinifera* proteins and four *P. trichocarpa* proteins group in the lignification-related R2R3-MYB clade. The *P. trichocarpa* genes encoding these proteins are located on LG_1 (*PtrMYB002 and PtrMYB003*) and on LG_IX (*PtrMYB020 and PtrMYB021*) in regions that are thought to be the paralogous product of the recent salicoid whole genome duplication event (Tuskan et al., 2006). The *P. trichocarpa* proteins are most similar to *EgMYB2* and to the *V. vinifera* members of this clade. Three of the four *Populus* R2R3-MYB genes exhibit high levels of transcript accumulation in xylem tissue (Fig. 5b), suggesting that function in this tissue has been retained for most family members since the duplication event. The transcript abundance profile also suggests that, like their counterparts in other plant species, these transcription factors also function in xylem-based processes, perhaps also regulating genes encoding enzymes of the lignin biosynthetic pathway. The retention of these apparent redundant functions in *Populus* speaks to the importance
of regulating the elaboration of diverse xylem cells in a woody angiosperm, and the necessity of the fine control of lignin biosynthesis associated therewith (Karpinska et al., 2004).

R2R3-MYB proteins are also well known regulators of anthocyanin biosynthesis in fruit, flowers and leaves. PRODUCTION OF ANTHOCYANIN PIGMENT (PAP) proteins, AtMYB75/PAP1 and AtMYB90/PAP2, are two of the best-characterised A. thaliana R2R3-MYB proteins involved in this process. AtMYB75/PAP1 is essential for both sucrose-induced (Teng et al., 2005) and light-induced anthocyanin accumulation (Cominelli et al., 2008). These genes form a clade (C27) that also include AtMYB113 and AtMYB114, which regulate late steps in anthocyanin biosynthesis in A. thaliana (Gonzalez et al., 2008), and Petunia x hybrida ANTHOCYANIN2 (PhAN2), which regulates anthocyanin in flowers (Quattrocchio et al., 1999) (Fig. 5c). Six Populus R2R3-MYB proteins grouped with this clade. Four of the genes encoding these proteins, PtrMYB116, PtrMYB117, PtrMYB118 and PtrMYB119 all had high levels of transcript accumulation in the male catkins (Fig. 5d). The male catkins of Populus are characteristically a deep red colour attributable to the abundance of anthocyanin-rich pollen grains. Anthocyanin pigments provide protection against damaging levels of UV radiation (Cominelli et al., 2008), and also function as precursors to conjugates that protect against pathogens and herbivory (Peters and Constabel, 2002). The apparently redundant activity of four R2R3-MYB proteins in male catkins underlines the importance of regulating the biosynthesis of these protectants in reproductive tissue. PtrMYB120 has the highest level of transcript accumulation in seedlings that have been grown in continuous darkness and then moved into the light for 3h, and may therefore play a role in invoking anthocyanin biosynthesis in vegetative tissues to protect them from the deleterious effects of UV light (Fig. 5d). PtrMYB113 shows low transcript accumulation throughout the tissues and organs examined, and may function to regulate anthocyanin biosynthesis in response to a stimulus that was not included in this study.

The genes encoding the proteins in the clade implicated in the regulation of anthocyanin biosynthesis have a distinctive arrangement the A. thaliana, V. vinifera, and P. trichocarpa genomes. In all three organisms, the genes are present as tandem duplications that appear to have arisen since they last shared a common ancestor (Fig. 5c, Supplemental Fig. S1). Notably, the R3 MYB repeat in P. trichocarpa and V. vinifera R2R3-MYB proteins in this clade is modified relative to the corresponding motif in the A. thaliana and Petunia x hybrida proteins. This modification takes the form of a four amino acid addition (QVkJQM) directly preceding the first conserved tryptophan in the R3 repeat (Supplemental Fig. S2a). These additional amino acids are encoded by sequences
present at the 3’ end of the second exon. Based on an alignment of the predicted mRNA molecules for the *A. thaliana*, *V. vinifera*, and *P. trichocarpa* genes in this clade, it does not appear that there has been an error in splice site prediction (Supplemental Fig. S2b). That is, the additional amino acids in the *Populus* R3 domain appear to be *bona fide* components of the motif. Given the role of this motif in protein-DNA interactions, the additional four amino acids in the *Populus* proteins could impact DNA binding. The fact that this motif is present in all six of the *Populus* members of the clade suggests that they are likely to bind to similar targets, but it remains to be determined if their binding-specificity or selectivity differs from their *A. thaliana* counterparts. Future studies could examine this possibility, and use it as the basis to explore co-evolution of DNA-binding domains relative to their cognate *cis*-acting DNA binding sites.

The high degree of redundancy within those clades where genes share highly similar transcript abundance profiles is consistent with the hypothesis that sub-functionalisation has occurred, with individual family members assuming non-redundant functions in the same tissue. In some cases, sub-functionalisation may have resulted in one of the MYB proteins acting as an activator of gene expression and another acting as a repressor. Co-expression of such antagonistic pairs of MYB proteins can produce a “gearing mechanism” where the regulation of shared target genes is a function of the relative abundance of a strong activator relative to a weaker one (Moyano et al., 1996). Alternatively, the high degree of redundancy within such clades may reflect the fact that there has not been adequate time for selection to purge one of the paralogues. It must be noted that genetic distance between paralogous pairs of R2R3-MYB genes was not a good predictor of the Pearson correlation coefficient describing their expression profiles (Supplemental Fig. S3). Namely, paralogous pairs separated by shorter genetic distances were no more likely to be co-expressed than were pairs separated by longer genetic distances.

**The *Populus* Electronic Fluorescent Pictograph browser facilitates rapid hypothesis testing related to MYB expression**

Identification of *Populus* R2R3-MYB genes with transcript abundance patterns or gene products that are phylogenetically close to MYB proteins with known function from other species, provides candidates for future studies aimed at thoroughly dissecting MYB function. Such studies would benefit from tools that enabled more intuitive representations of transcript abundance patterns from the PopGenExpress compendium, so that rapid comparisons could be made to putative orthologues from better-characterised species such as *A. thaliana*. To enable community-wide, simple, graphical
representation of PopGenExpress transcript abundance data, a web-based tool was devised, the Populus Electronic Fluorescent Pictograph (eFP) browser (Fig. 6) (http://www.bar.utoronto.ca/efppop/cgi-bin/efpWeb.cgi). The poplar eFP browser is based on eFP tools for A. thaliana and Mus musculus (Winter et al., 2007). For the poplar eFP browser, diagrams of poplar tissues, organs or growth conditions are shaded with colours corresponding to the quantity of transcript for a given gene under that condition.

The Arabidopsis eFP browser tool has already proven to be highly useful in the display and interpretation of A. thaliana transcript abundance data. Use of the comparable tool for Populus (Fig. 6) will also enable straightforward, intuitive representation of Populus transcript abundance data, as well as simple comparison of transcript abundance distribution for a given Populus gene against its putative A. thaliana counterpart. Comparison of eFP browser profiles for homologous Arabidopsis and Populus genes facilitates direct testing of the hypothesis that the homologues may respond to the same stimuli or play a role in the same processes in their respective species. This simple analysis can be extended to any pair of potentially orthologous genes, across the entire set of genes probed by the Affymetrix array. To facilitate such hypothesis testing and development, all A. thaliana and P. trichocarpa orthologues have been pre-computed using the OrthoMCL algorithm and are displayed as links in the eFP browser.

CONCLUSIONS

Detailed annotation and phylogenetic analysis of the entire complement of Populus trichocarpa R2R3-MYB genes and their protein products reveals the striking diversification that has occurred in this gene family in Populus. Some of this diversification is attributable to genome duplication, but unequal expansion in particular clades, and the manifestation of entire clades lacking A. thaliana family members suggests that some of the diversification in Populus R2R3-MYB family members may contribute to lineage-specific phenotypic innovations. Here we have identified some of the family members that may contribute to these innovations, through an examination of both the R2R3-MYB phylogeny and patterns of transcript abundance. These genes will form the basis for future hypothesis tests involving gain-of-function and loss-of-function studies aimed at clarifying their roles in Populus growth, development and survival.

We also provide a collection of data and an associated bioinformatics tool that enable the research community to examine, visualise and formulate hypotheses based on transcript abundance of
R2R3-MYB-encoding genes, or, indeed, any other gene in *Populus* that is represented on the Affymetrix Poplar GeneChip. By pre-computing *A. thaliana* – *Populus* orthologues, we also enable the easy cross-species expression viewing to identify the true functional orthologue in *Populus* of a given *A. thaliana* gene, if it exists. The gene expression database is expandable, and will acquire greater value as other researchers add their GeneChip data to the core compendium we have established. This will add greatly to the expanding toolbox available to characterise the molecular mechanisms underpinning the basic biology of an ecologically dominant, and economically important tree genus.

**MATERIALS AND METHODS**

**Plant Material**

Plant material was collected from clonally-propagated, 8-week-old *Populus balsamifera* saplings (Clone 1006, Alberta Pacific Forest Industries Inc., Boyle, AB, Canada) grown in a climate-controlled growth chamber (mature leaf, young leaf, roots, differentiating xylem), or from *P. balsamifera* seeds (etiolated light and dark seedlings) (Seedlot number 20071015, National Tree Seed Centre, Fredericton, NB, Canada), or from mature *P. balsamifera* trees grown by Alberta Pacific Forest Industries Inc. in a field trial in Grassland, AB, Canada (male and female catkins). Trees established in the growth chamber were grown in Sunshine Mix (Sun Gro Horticulture, Vancouver, BC, Canada) with a 16h photoperiod, a maximum daytime temperature of 22°C and a minimum nighttime temperature of 17°C. All tissues were collected 8h into the light phase during a 16h:8h light:dark cycle, except for the midnight samples, which were collected at 4h after the shift from light to dark in the same cycle. All tissues were collected from saplings without water limitations. For the comparison of dark-grown seedlings *versus* those grown in the dark and then exposed to light for 3h, seeds were germinated in a dark, in a growth chamber on wetted filter paper in petri plates with a constant temperature of 21°C, and were grown in this way for five days. On the sixth day, half of the plates were exposed to light (150μmol) for three hours. At this time seedlings were collected from plates exposed to light and from plates in continuous darkness. All collected plant materials were flash frozen in liquid nitrogen upon harvesting and stored at -80°C until RNA was extracted. Each sample comprises pooled material from three individuals, except for the seedling samples, which comprise pooled material from 20 individuals. Mature leaf samples comprise the first fully expanded leaf of three saplings, including the petiole. Young leaf samples include the first leaf, with its petiole, that was completely uncurled, but was still shiny and much smaller than the mature
leaves. Root samples include the distal 15 cm of well-rinsed roots mass. Differentiating xylem was collected from the bottom 6 inches of the stem by peeling off the bark and immediately scraping the surface of the exposed wood into micro collection tubes containing liquid nitrogen. Seedlings were grown for 5 days in the dark, on the sixth day, one half of the seedlings were transferred to the light and samples (comprising entire seedlings) were collected at 30 minutes after transfer to light. Male and female catkins were collected on April 30, 2007.

Phylogenetic Analysis

*A. thaliana* R2R3- and 3R-MYB gene identifiers were obtained from Stracke et al. (2001) and the corresponding protein sequences were downloaded from The Arabidopsis Information Resource (TAIR, http://arabidopsis.org) (Rhee et al., 2003) (Supplemental Table S3). Different identifiers have been used for AtMYB8 (At1g35515), AtMYB39 (At4g17785), and AtMYB118 (At3g27785) in this paper than in Stracke et al. (2001).

Gene identifiers for 94 *V. vinifera* R2R3-MYB were obtained from Matus et al. (2008) and the corresponding protein sequences were downloaded from the International Grape Genome Program's (IGGP) website (Program, 2008). It is important to note that, while Matus et al. (2008) identified 108 putative R2R3-MYB-encoding genes, careful scrutiny of the coding sequences revealed that 15 of these genes actually represent single repeat MYB proteins, and so were excluded from the analysis described herein. An additional 25 R2R3- and five 3R-MYB genes were identified on the IGGP's website using the search terms "R2R3" and "MYB transcription factor" (Supplemental Table S4).

*P. trichocarpa* MYB gene models were retrieved from the JGI *Populus trichocarpa* v1.1 JGI website (Joint Genome Institute, http://genome.jgi-psf.org/Potr1_1/Potr1_1.home.html) (Institute, 2008). Putative MYB genes in the *P. trichocarpa* genome were identified using the PAC v0.1 Ortholog Finder using *A. thaliana* R2R3-MYBs as reference (Supplemental Table S1). MYB gene models were identified using the KOG keyword MYB. All *P. trichocarpa* and *V. vinifera* models were manually inspected to ensure that the putative gene models contained two or three MYB DBDs, and that the gene models mapped to unique loci in their respective genomes. It is important to note that a proportion of the 21 *P. trichocarpa* R2R3-MYB genes that have not been assigned to a linkage group may be alleles of genes already placed on a linkage group, but as this has not yet been resolved empirically, these 21 genes were treated as distinct loci.
An additional 22 R2R3-MYB transcription factors from other tree species were included in the phylogeny - 7 from *Pinus taeda*, 13 from *Picea glauca* and two from *Eucalyptus gunnii* (Supplemental Table S5). A further 32 'landmark' MYB proteins from a variety of other organisms were also included in the analysis (Supplemental Table S6) (Jiang et al., 2004). Sequences were retrieved from NCBI protein database (NCBI Entrez, http://www.ncbi.nlm.nih.gov/sites/entrez).

Phylogenetic analysis included the above 506 R2R3-MYB proteins and 15 3R-MYB proteins. The full-length amino acid sequences were aligned using MAFFT using the G-INS-i algorithm (Katoh et al., 2005). A neighbour-joining tree was constructed using MEGA 4 (Tamura et al., 2007) with settings for Jones-Taylor-Thornton (JTT) substitution model and a Gamma parameter of 1.0 to allow for the uneven rates of substitution across the length of the protein, including the highly conserved amino terminal DNA binding domain and the more divergent carboxy-terminal activation domain. Pairwise gap deletion was used to ensure that the carboxy-terminal residues, which are more divergent in MYB proteins, could contribute to the topology of the neighbour-joining tree.

**RNA Extraction and Microarray Hybridisation**

Plant material was ground to a fine powder under liquid nitrogen, and total RNA was extracted from each sample using the procedure described by Chang *et al.* (Chang et al., 1993) for all materials except the seedling samples, which were extracted using the Trizol method (Invitrogen, Burlington, ON, Canada). RNA quality was determined electrophoretically. For each sample, 10 μg of total RNA was reverse transcribed, labeled, and hybridised to the Poplar Genome Array according to manufacturer's protocols (Affymetrix, Santa Clara, CA) at the Centre for Applied Genome Evolution & Function at the University of Toronto. The Poplar Genome Array includes 61,251 probe sets representing more than 56,055 transcripts (Affymetrix, Santa Clara, CA). The GeneChip probe design is based on the Joint Genome Institute's *Populus trichocarpa* genome project's predicted gene set v1.1 and all publicly available EST and mRNA sequences for all *Populus* species available through UniGene Build #6 and GenBank in spring 2005. The array probes for transcripts derived from 13 *Populus* species in addition to the fully sequenced *P. trichocarpa* genome. For each condition, RNA from three replicate biological samples was extracted and hybridized to a Poplar GeneChip.
Microarray Analysis

GeneChip data analysis was performed using the BioConductor suite (Gentleman et al., 2004) in R (R: A language and environment for statistical computing, http://www.R-project.org) (R Core Development Team, 2007) using the affy package (Gautier et al., 2004). Pre-processing of the arrays was done using GC-robust multiarray analysis (GCRMA) (Gentleman et al., 2004). Probe sets corresponding to the putative R2R3-MYB and 3R-MYB transcription factors were identified using the Probe Match tool on the NetAffx Analysis Centre website (Table 1) (Netaffx, http://www.affymetrix.com/analysis/index.affx). The transcript abundance levels for probe sets corresponding to the putative MYB genes in the *P. trichocarpa* genome were extracted from the complete data set and the probe sets were then clustered using hierarchical clustering based on Pearson correlation coefficients (Ploner, 2008).

Poplar eFP Browser

A Poplar eFP Browser (http://www.bar.utoronto.ca/efppop/cgi-bin/efpWeb.cgi) was developed based on the Bio-Array Resource database framework described in Toufighi et al. (Toufighi et al., 2005) and the eFP engine developed by Winter et al. (Winter et al., 2007). Briefly, database tables were generated using MySQL to archive meta-information for the samples used to produce the transcriptome data in this study, as well as to store the expression data themselves, after the MIAME convention (Brazma et al., 2001). This database may be queried through other tools described in Toufighi et al. (2005), such as Expression Browser, Project Browser, and Expression Angler for identifying co-expressed genes of interest (Toufighi et al., 2005). In the case of the poplar gene expression database, we normalized the Affymetrix data using the GCOS algorithm, with a TGT value of 500. Diagrammatic representations of poplar were generated and processed as described in Winter et al. (Winter et al., 2007) to create an input for the eFP Browser. In addition, an XML control file was created to instruct the eFP engine for proper functioning. Finally, *A. thaliana* – *Populus* orthologues based on the *P. trichocarpa* gene models were computed using the OrthoMCL algorithm (Li et al., 2003). These were stored in a separate database along with pre-computed *A. thaliana* – *Populus* sequence alignments, generated using MAFFT using the G-INS-I algorithm (Katoh et al., 2005) and displayed on demand using Nigel Brown’s MView (Brown et al., 1998) program.
ACKNOWLEDGEMENTS

We are most grateful to Dr. Levi Waldron and Ms. Than Nguyen for superb technical assistance, to Ms. Josephine McKeever for excellent renderings of plant organs, and to Dr. Barb Thomas and Mr. David Kamelchuk for generously providing *Populus* plant material. *Populus* seeds were generously provided by the Canadian National Tree Seed Centre (Fredericton, NB, Canada).

LITERATURE CITED

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15**: 63-78

Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D, Shinozaki K (1997) Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Plant Cell* **9**: 1859-1868

Baumann K, Perez-Rodriguez M, Bradley D, Venail J, Bailey P, Jin H, Koes R, Roberts K, Martin C (2007) Control of cell and petal morphogenesis by R2R3 MYB transcription factors. *Development* **134**: 1691-1701

Bedon F, Grima-Pettenati J, Mackay J (2007) Conifer R2R3-MYB transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). *BMC Plant Biol* **7**: 17

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. *Annu Rev Plant Biol* **54**: 519-546

Bradshaw HD, Ceulemans R, Davis J, Stettler R (2000) Emerging model systems in plant biology: *Poplar (Populus)* as a model forest tree. *Journal of Plant Growth Regulation* **19**: 306-313

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matses JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat Genet* **29**: 365-371

Brown NP, Leroy C, Sander C (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**: 380-381

Brunner AM, Busov VB, Strauss SH (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends in Plant Science* **9**: 49-56
Chang SJ, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter 11: 4

Chen YH, Yang XY, He K, Liu MH, Li JG, Gao ZF, Lin ZQ, Zhang YF, Wang XX, Qiu XM, Shen YP, Zhang L, Deng XH, Luo JC, Deng XW, Chen ZL, Gu HY, Qu LJ (2006) The MYB transcription factor superfamily of arabidopsis: Expression analysis and phylogenetic comparison with the rice MYB family. Plant Molecular Biology 60: 107-124

Cominelli E, Gusmaroli G, Allegra D, Galbiati M, Wade HK, Jenkins GI, Tonelli C (2008) Expression analysis of anthocyanin regulatory genes in response to different light qualities in Arabidopsis thaliana. Journal of Plant Physiology 165: 886-894

Deluc L, Barrieu F, Marchive C, Lauvergeat V, Decendit A, Richard T, Carde JP, Merillon JM, Hamdi S (2006) Characterization of a grapevine R2R3-MYB transcription factor that regulates the phenylpropanoid pathway. Plant Physiol 140: 499-511

Doebly J, Lukens L (1998) Transcriptional regulators and the evolution of plant form. Plant Cell 10: 1075-1082

Durbarry A, Vizir I, Twell D (2005) Male germ line development in Arabidopsis. duo pollen mutants reveal gametophytic regulators of generative cell cycle progression. Plant Physiol 137: 297-307

Feng CP, Andreasson E, Maslak A, Mock HP, Mattsson O, Mundy J (2004) Arabidopsis MYB68 in development and responses to environmental cues. Plant Science 167: 1099-1107

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307-315

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80

Gocal GF, Sheldon CC, Gubler F, Moritz T, Bagnall DJ, MacMillan CP, Li SF, Parish RW, Dennis ES, Weigel D, King RW (2001) GAMYB-like genes, flowering, and gibberellin signaling in Arabidopsis. Plant Physiol 127: 1682-1693

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J (2005) EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary cell wall formation and lignin biosynthesis. Plant J 43: 553-567
Gonzalez A, Zhao M, Leavitt JM, Lloyd AM (2008) Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. Plant J 53: 814-827

Haga N, Kato K, Murase M, Araki S, Kubo M, Demura T, Suzuki K, Muller I, Voss U, Jurgens G, Ito M (2007) R1R2R3-Myb proteins positively regulate cytokinesis through activation of KNOLLE transcription in Arabidopsis thaliana. Development 134: 1101-1110

Hartmann U, Sagasser M, Mehrtens F, Stracke R, Weisshaar B (2005) Differential combinatorial interactions of cis-acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes. Plant Mol Biol 57: 155-171

Higginson T, Li SF, Parish RW (2003) AtMYB103 regulates tapetum and trichome development in Arabidopsis thaliana. Plant J 35: 177-192

International Grape Genome Program (2008) http://www.genoscope.cns.fr/externe/English/Projets/Projet_ML/projet.html. In. Genoscope

Ito M (2005) Conservation and diversification of three-repeat Myb transcription factors in plants. J Plant Res 118: 61-69

Ito M, Araki S, Matsunaga S, Itoh T, Nishihama R, Machida Y, Doonan JH, Watanabe A (2001) G2/M-phase-specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors. Plant Cell 13: 1891-1905

Jiang C, Gu X, Peterson T (2004) Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of Arabidopsis and Oryza sativa L. ssp. indica. Genome Biol 5: R46

Jin H, Cominelli E, Bailey P, Parr A, Mehrtens F, Jones J, Tonelli C, Weisshaar B, Martin C (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. EMBO J 19: 6150-6161

Jin H, Martin C (1999) Multifunctionality and diversity within the plant MYB-gene family. Plant Mol Biol 41: 577-585

Joint Genome Institute (2008) Populus trichocarpa v1.1. DOE Joint Genome Institute

Jung C, Seo JS, Han SW, Koo YJ, Kim CH, Song SI, Nahm BH, Choi YD, Cheong JJ (2008) Overexpression of AtMYB44 enhances stomatal closure to confer abiotic stress tolerance in transgenic Arabidopsis. Plant Physiol 146: 623-635

Karpinska B, Karlsson M, Srivastava M, Stenberg A, Schrader J, Sterky F, Bhalerao R, Wingsle G (2004) MYB transcription factors are differentially expressed and regulated during secondary vascular tissue development in hybrid aspen. Plant Mol Biol 56: 255-270
Kasahara RD, Portereiko MF, Sandaklie-Nikolova L, Rabiger DS, Drews GN (2005) MYB98 is required for pollen tube guidance and synergid cell differentiation in Arabidopsis. Plant Cell 17: 2981-2992

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518

Keller T, Abbott J, Moritz T, Doerner P (2006) Arabidopsis REGULATOR OF AXILLARY MERISTEMS1 controls a leaf axil stem cell niche and modulates vegetative development. Plant Cell 18: 598-611

Kranz H, Scholz K, Weisshaar B (2000) c-MYB oncogene-like genes encoding three MYB repeats occur in all major plant lineages. Plant J 21: 231-235

Kranz HD, Denekamp M, Greco R, Jin H, Leyva A, Meissner RC, Petroni K, Urzainqui A, Bevan M, Martin C, Smeekens S, Tonelli C, Paz-Ares J, Weisshaar B (1998) Towards functional characterisation of the members of the R2R3-MYB gene family from Arabidopsis thaliana. Plant J 16: 263-276

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189

Lipsick JS (1996) One billion years of Myb. Oncogene 13: 223-235

Matus JT, Aquea F, Arce-Johnson P (2008) Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. BMC Plant Biol 8: 83

Mehrtens F, Kranz H, Bednarek P, Weisshaar B (2005) The Arabidopsis transcription factor MYB12 is a flavonol-specific regulator of phenylpropanoid biosynthesis. Plant Physiol 138: 1083-1096

Meissner RC, Jin H, Cominelli E, Denekamp M, Fuertes A, Greco R, Kranz HD, Penfield S, Petroni K, Urzainqui A, Martin C, Paz-Ares J, Smeekens S, Tonelli C, Weisshaar B, Baumann E, Klimyk V, Marillonnet S, Patel K, Speulman E, Tissier AF, Bouchez D, Jones JJ, Pereira A, Wisman E, et al. (1999) Function search in a large transcription factor gene family in Arabidopsis: assessing the potential of reverse genetics to identify insertion mutations in R2R3 MYB genes. Plant Cell 11: 1827-1840

Miranda M, Ralph SG, Mellway R, White R, Heath MC, Bohlmann J, Constabel CP (2007) The transcriptional response of hybrid poplar (Populus trichocarpa x P-deltoides) to infection by Melampsora medusae leaf rust involves induction of flavonoid pathway genes leading to the accumulation of proanthocyanidins. Molecular Plant-Microbe Interactions 20: 816-831
Moyano E, Martinez-Garcia JF, Martin C (1996) Apparent redundancy in myb gene function provides gearing for the control of flavonoid biosynthesis in antirrhinum flowers. Plant Cell 8: 1519-1532

Muller D, Schmitz G, Theres K (2006) Blind homologous R2R3 Myb genes control the pattern of lateral meristem initiation in Arabidopsis. Plant Cell 18: 586-597

Ogata K, Morikawa S, Nakamura H, Hojo H, Yoshimura S, Zhang R, Aimoto S, Ametani Y, Hirata Z, Sarai A, et al. (1995) Comparison of the free and DNA-complexed forms of the DNA-binding domain from c-Myb. Nat Struct Biol 2: 309-320

Okada M, Akimaru H, Hou DX, Takahashi T, Ishii S (2002) Myb controls G(2)/M progression by inducing cyclin B expression in the Drosophila eye imaginal disc. EMBO J 21: 675-684

Oppenheimer DG, Herman PL, Sivakumaran S, Esch J, Marks MD (1991) A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. Cell 67: 483-493

Patzlaff A, Mclnnis S, Courtenay A, Surman C, Newman LJ, Smith C, Bevan MW, Mansfield S, Whetten RW, Sederoff RR, Campbell MM (2003) Characterisation of a pine MYB that regulates lignification. Plant J 36: 743-754

Perez-Rodriguez M, Jaffe FW, Butelli E, Glover BJ, Martin C (2005) Development of three different cell types is associated with the activity of a specific MYB transcription factor in the ventral petal of Antirrhinum majus flowers. Development 132: 359-370

Peters DJ, Constabel CP (2002) Molecular analysis of herbivore-induced condensed tannin synthesis: cloning and expression of dihydroflavonol reductase from trembling aspen (Populus tremuloides). Plant J 32: 701-712

Ploner A (2008) Heatplus.1.10.0. In Bioconductor,

Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, Koes R (1999) Molecular analysis of the anthocyanin2 gene of Petunia and its role in the evolution of flower color. Plant Cell 11: 12

R Core Development Team (2007) R: A language and Environment for Statistical Computing www.R-project.org. In. R Foundation for Statistical Computing, Vienna, Austria

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 31: 224-228

Riechmann JL, Heard J, Martin G, Reuber L, Jiang CZ, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman
BK, Yu CL (2000) Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. Science 290: 2105-2110

Rosinski JA, Atchley WR (1998) Molecular evolution of the Myb family of transcription factors: Evidence for polyphyletic origin. Journal of Molecular Evolution 46: 74-83

Schmitz G, Tillmann E, Carriero F, Fiore C, Cellini F, Theres K (2002) The tomato Blind gene encodes a MYB transcription factor that controls the formation of lateral meristems. Proc Natl Acad Sci U S A 99: 1064-1069

Shin B, Choi G, Yi H, Yang S, Cho I, Kim J, Lee S, Paek NC, Kim JH, Song PS (2002) AtMYB21, a gene encoding a flower-specific transcription factor, is regulated by COP1. Plant J 30: 23-32

Stracke R, Ishihara H, Huep G, Barsch A, Mehrtens F, Niehaus K, Weisshaar B (2007) Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling. Plant J 50: 660-677

Stracke R, Werber M, Weisshaar B (2001) The R2R3-MYB gene family in Arabidopsis thaliana. Curr Opin Plant Biol 4: 447-456

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599

Teng S, Keurentjes J, Bentsink L, Koornneef M, Smeekens S (2005) Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. Plant Physiol 139: 1840-1852

Tominaga R, Iwata M, Okada K, Wada T (2007) Functional analysis of the epidermal-specific MYB genes CAPRICE and WEREWOLF in Arabidopsis. Plant Cell 19: 2264-2277

Toufighi K, Brady SM, Austin R, Ly E, Prohart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J 43: 153-163

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blauze D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrikkson R, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F,
Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryabo V, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**: 1596-1604

Urao T, Noji M, Yamaguchi-Shinozaki K, Shinozaki K (1996) A transcriptional activation domain of *At*MYB2, a drought-inducible *Arabidopsis* Myb-related protein. Plant J **10**: 1145-1148

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Ozyerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematte L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Saramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**: e1326

Wada T, Tachibana T, Shimura Y, Okada K (1997) Epidermal cell differentiation in Arabidopsis determined by a Myb homolog, CPC. Science **277**: 1113-1116

Weatherbee SD, Halder G, Kim J, Hudson A, Carroll S (1998) Ultrabithorax regulates genes at several levels of the wing-patterning hierarchy to shape the development of the *Drosophila* haltere. Genes & Development **12**: 1474-1482

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provant NJ (2007) An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* **2**: e718

Yang XY, Li JG, Pei M, Gu H, Chen ZL, Qu LJ (2007) Over-expression of a flower-specific transcription factor gene *AtMYB24* causes aberrant anther development. Plant Cell Rep **26**: 219-228

Zhang ZB, Zhu J, Gao JF, Wang C, Li H, Zhang HQ, Zhang S, Wang DM, Wang QX, Huang H, Xia HJ, Yang ZN (2007) Transcription factor *AtMYB103* is required for anther development by regulating tapetum development, callose dissolution and exine formation in *Arabidopsis*. Plant J **52**: 528-538
FIGURE LEGENDS

Figure 1: The R2 and R3 MYB repeats are highly conserved across all R2R3-MYB proteins in the *Populus trichocarpa* genome. The sequence logos of the (a) R2 and (b) R3 MYB repeats are based on full-length alignments of all *Ptr*R2R3-MYB proteins. The bit score indicates the information content for each position in the sequence. Highly conserved tryptophan residues required for DNA binding are highlighted in red. *Ptr*MYB proteins in C27 have an additional four amino acids directly before the first conserved tryptophan in R3 (Supplementary Figure 2a).

Figure 2: Phylogenetic relationships and subgroup designations in MYB proteins from *Populus trichocarpa*, *Vitis vinifera*, and *Arabidopsis thaliana*. The neighbour-joining tree includes 192 R2R3-MYB proteins from *Populus*, 118 from *Vitis* and 126 from *Arabidopsis*, and a further 54 from other plant species, in addition to five 3R-MYB proteins from each of *Populus*, *Vitis*, and *Arabidopsis*. The proteins are clustered into 49 subgroups (triangles), designated with a subgroup number (eg. C1). Eleven proteins did not fit well into clusters (lines). Some ‘landmark’ R2R3-MYB proteins are listed to the right of the clades for reference. The membership of each subgroup is described the table (b) The scale bar under the tree represents 0.5 substitutions. Several clades are highlighted that exemplify lineage specific R2R3-MYB gene expansion. C3 and C25 (green) show dramatic expansion in the *P. trichocarpa* lineage. C13 and C37 (yellow) do not include any *Ptr*MYB proteins. C14 and C22 (blue) show limited expansion in the *P. trichocarpa* lineage. Many clades do not include any *A. thaliana* R2R3-MYB proteins (orange). The uncompressed tree with full taxa names is available as Supplementary Figure 1.

Figure 3: R2R3-MYB and 3R-MYB proteins are present on all linkage groups (LG) in the *P. trichocarpa* genome. One hundred and seventy two of the *Populus* R2R3-MYB genes are mapped to LG according to JGI Poplar Genome v1.1. Each vertical bar represents one LG drawn to scale. The identities of the MYB genes are indicated to the right of the linkage group on which they are located with horizontal lines intersecting with the LG indicating their position. In cases where multiple MYB genes are located in close proximity on a LG, lines connecting individual gene names and the LG have been omitted. In these cases curved lines have been used to bracket the relevant gene names and to display their genomic location. The coloured boxes indicate groups of MYB proteins.
with paralogous and syntenic genes on two LG. The groups indicated are not exhaustive, but serve to indicate the widespread evidence of the Salicoid-specific whole genome duplication event in the *P. trichocarpa* genome.

**Figure 4:** *Populus* R2R3-MYB genes exhibit differential expression across a range of tissues, organs, and treatments. The patterns of relative transcript accumulation of each of 180 R2R3-MYB genes and five 3R-MYB genes as determined by microarray analysis is presented as a heatmap with red indicating higher levels and green indicating lower levels of transcript accumulation. Each column represents a discreet biological sample and all treatments are presented as biological triplicate replicates (CL = seedlings grown in continuous light, DL = seedlings grown in continuous darkness, and then transferred to light for 3h, CD = seedlings grown in continuous darkness, YL = young leaf, ML = mature leaf, R = root, DX = differentiating xylem, FC = female catkins, MC = male catkins). Data are normalised within each row.

**Figure 5:** *Ptr*R2R3-MYB genes show conserved patterns of transcript accumulation across tissues. (a) Clade 10 (C10) includes R2R3-MYB genes isolated from the wood forming tissues of a number of species including *Picea glauca* (*Pg*), *Pinus taeda* (*Pt*), *Eucalyptus gunnii* (*Eg*), *Vitis vinifera* (*Vv*), and *A. thaliana* (*At*) as well as four previously uncharacterized R2R3-MYB genes from *P. trichocarpa* (*Ptr*). (b) The relative transcript accumulation of each of the *Ptr*R2R3-MYB genes genes in C10 as determined by microarray analysis. (c) Clade 27 (C27) includes several well-known R3R3-MYB genes with demonstrated roles in anthocyanin biosynthesis in a variety of organisms including *AtPAP1/MYB75*, *AtPAP2/MYB90* and *Petunia x hybrida* Anthocyanin 2 (*PhAN2*) and six previously uncharacterised *Ptr*R2R3-MYB genes. (d) The relative transcript accumulation of each of the *Ptr*R2R3-MYB genes genes in C27 as determined by microarray analysis. The scale bars under the trees represent 0.2 substitutions. The heatmaps are clustered by Pearson correlation coefficient. Red indicates higher levels and green represents lower levels of transcript accumulation.
Figure 6: Electronic fluorescence pictograph (eFP) display transcript accumulation patterns across a variety of Populus organs and treatments. Poplar eFP browser presents transcript accumulation pattern of *Ptr*MYB025 in a variety of tissues and organs. In all cases, red colouring indicates higher levels of transcript accumulation and yellow indicates a lower level of transcript accumulation.

TABLES

Table 1: R2R3- and 3R-MYB gene complement of *A. thaliana*, *V. vinifera* and *P. trichocarpa*.

| Organism       | R2R3-MYB | 3R-MYB |
|----------------|----------|--------|
| *A. thaliana*   | 126      | 5      |
| *V. vinifera*   | 118      | 5      |
| *P. trichocarpa*| 192      | 5      |

SUPPLEMENTAL DATA

Supplemental Figure S1: Phylogenetic relationships and subgroup designations in MYB proteins in *Populus*, *Vitis*, and *Arabidopsis*.

Supplemental Figure S2: *Ptr*R2R3-MYB genes in the Clade 27 have a modified R3 DNA binding domain. (a) Amino acid sequence logo corresponding to the modified region, showing the location of 4 additional residues found in Clade 27 members relative to the remainder of the *Ptr*R2R3-MYB family. (b) DNA sequence alignment showing that the nucleotides encoding the 4 additional amino acids are located immediately upstream of the 5’ splice donor site of intron 2.

Supplemental Figure S3: Pairwise Pearson correlation coefficients of the expression profiles of 56 paralogous R2R3-MYB gene pairs as a function of the genetic distance separating the pair. No evidence of a relationship between these factors was observed. Paralogous pairs were included only if both genes had corresponding probe sets on the array.

Supplemental Table S1: R2R3- and 3R-MYB encoding genes from *Populus trichocarpa*. When a MYB number appears to be absent, this has arisen from the fact that, when the *Populus* genome was annotated, some gene models were mis-annotated as R2R3-MYB-encoding genes, when they
were, in fact, 1R-MYB-encoding sequences. As was the case with the *Vitis* 1R-MYB sequences, the *Populus* 1R-MYB sequences were excluded from the construction of the R2R3-MYB and 3R-MYB phylogeny.

**Supplemental Table S2:** Membership of subgroups in the *P. trichocarpa*, *A. thaliana*, and *V. vinifera* phylogenetic tree.

**Supplemental Table S3:** R2R3- and 3R-MYB encoding genes from *Arabidopsis thaliana*.

**Supplemental Table S4:** R2R3- and 3R-MYB encoding genes from *Vitis vinifera*. ¹Indicates gene model identified by Matus et al. (2008).

**Supplemental Table S5:** R2R3- and 3R-MYB encoding genes from a variety of tree species.

**Supplemental Table S6:** Landmark R2R3- and 3R-MYB encoding genes from a variety of plant species.
Figure 1: The R2 and R3 MYB repeats are highly conserved across all R2R3-MYB proteins in the *Populus trichocarpa* genome. The sequence logos of the (a) R2 and (b) R3 MYB repeats are based on full-length alignments of all *PtrR2R3-MYB* proteins. The bit score indicates the information content for each position in the sequence. Highly conserved tryptophan residues required for DNA binding are highlighted in red. *PtrMYB* proteins in C27 have an additional four amino acids directly before the first conserved tryptophan in R3 (Supplementary Figure 2a).
Figure 2: Phylogenetic relationships and subgroup designations in MYB proteins from *Populus trichocarpa*, *Vitis vinifera*, and *Arabidopsis thaliana*. The neighbour-joining tree includes 192 R2R3-MYB proteins from *Populus*, 119 from *Vitis* and 126 from *Arabidopsis*, and a further 54 from other plant species, in addition to five 3R-MYB proteins from each of *Populus*, *Vitis*, and *Arabidopsis*. The proteins are clustered into 49 subgroups (triangles), designated with a subgroup number (eg. C1). Eleven proteins did not fit well into clusters (lines). Some ‘landmark’ R2R3-MYB proteins are listed to the right of the clades for reference. The membership of each subgroup is described in table (b). The scale bar under the tree represents 0.5 substitutions. Several clades are highlighted that exemplify lineage specific R2R3-MYB gene expansion. C3 and C25 (green) show dramatic expansion in the *P. trichocarpa* lineage. C13 and C37 (yellow) do not include any *PtrMYB* proteins. C14 and C22 (blue) show limited expansion in the *P. trichocarpa* lineage. Many clades do not include any *A. thaliana* R2R3-MYB proteins (orange). The uncompressed tree can be downloaded from the Supplementary Material.
Figure 3: R2R3-MYB and 3R-MYB proteins are present on all linkage groups (LG) in the *P. trichocarpa* genome. One hundred and seventy two of the *Populus* R2R3-MYB genes are mapped to LG according to JGI Poplar Genome v1.1. Each vertical bar represents one LG drawn to scale. The identities of the MYB genes are indicated to the right of the linkage group on which they are located with horizontal lines intersecting with the LG indicating their position. In cases where multiple MYB genes are located in close proximity on a LG, lines connecting individual gene names and the LG have been omitted. In these cases curved lines have been used to bracket the relevant gene names and to display their genomic location. The coloured boxes indicate groups of MYB proteins with paralogous and syntenic genes on two LG. The groups indicated are not exhaustive, but serve to indicate the widespread evidence of the salicoid-specific whole genome duplication event in the *P. trichocarpa* genome.
Figure 4: *Populus* R2R3-MYB genes exhibit differential expression across a range of tissues, organs, and treatments. The patterns of relative transcript accumulation of each of 180 R2R3-MYB genes and five 3R-MYB genes as determined by microarray analysis is presented as a heatmap with red indicating higher levels and green indicating lower levels of transcript accumulation. Each column represents a discreet biological sample and all treatments are presented as biological triplicate replicates (CL = seedlings grown in continuous light, DL = seedlings grown in continuous darkness, and then transferred to light for 3h, CD = seedlings grown in continuous darkness, YL = young leaf, ML = mature leaf, R = root, DX = differentiating xylem, FC = female catkins, MC = male catkins). Data are normalised within each row.
Figure 5: *PtrR2R3-MYB* genes show conserved patterns of transcript accumulation across tissues. (a) Clade 10 (C10) includes R2R3-MYB genes isolated from the wood forming tissues of a number of species including *Picea glauca* (Pg), *Pinus taeda* (Pt), *Eucalyptus gunnii* (Eg), *Vitis vinifera* (Vv), and *A. thaliana* (At) as well as four previously uncharacterized R2R3-MYB genes from *P. trichocarpa* (Ptr). (b) The relative transcript accumulation of each of the *PtrR2R3-MYB* genes in C10 as determined by microarray analysis. (c) Clade 27 (C27) includes several well-known R3R3-MYB genes with demonstrated roles in anthocyanin biosynthesis in a variety of organisms including AtPAP1/MYB75, AtPAP2/MYB90 and *Petunia x hybrida* Anthocyanin 2 (PhAN2) and six previously uncharacterised *PtrR2R3-MYB* genes. (d) The relative transcript accumulation of each of the *PtrR2R3-MYB* genes in C27 as determined by microarray analysis. The scale bars under the trees represent 0.2 substitutions. The heatmaps are clustered by Pearson correlation coefficient. Red indicates higher levels and green represents lower levels of transcript accumulation.
Figure 6: Electronic fluorescence pictograph (eFP) display transcript accumulation patterns across a variety of Populus organs and treatments. Poplar eFP browser (a) presents transcript accumulation pattern of *PtrMYB025* in a variety of tissues and organs. In all cases, red colouring indicates higher levels of transcript accumulation and yellow indicates a lower level of transcript accumulation.