Natural variation in the serially duplicated *Production of Anthocyanin Pigment* loci and anthocyanin accumulation in *Arabidopsis thaliana* (Brassicaceae)

A thesis presented in partial fulfilment of the requirements for the Degree of Masters of Science in Plant Biology at Massey University, Palmerston North, New Zealand

Matthew Butcher

2013

I dedicate this thesis to my future children.

May they have a rich father and a beautiful mother.

Acknowledgements

I would like to thank Dr. Vaughan Symonds for his guidance, supervision and advice as well as providing the facilities to complete this work. Without his immeasurable patience and understanding throughout the course of my project I would not have made it through to the end.

I also want to thank past and present members of the LoSTLab, in no particular order, Dr. Jen Tate, Tina, Rowan, Nick, Jill, Jessie, Amir, Fronny, Cindy, Megan and Kay for making it a fun, constructive and positive environment to work in. I'd especially like to thank fellow LostLab member Rebecca Bloomer for forging the way before me, her never-ending knowledge of all things developmental genetics and, most importantly, listening to me moan and complain the whole way through.

Thanks to all my friends, especially Todd, Sam and Leigh, for their unrelenting support as well as the welcome distractions at the right times which kept me sane these past few years.

Last, though certainly not least, I would like to thank my parents Steve and Gail, my brother Nathaniel, my sister Hannah, and the rest of my family for their support in the past and their continued support looking to the future. I couldn't have done this without them.

Contents

Figures

Tables

1. Abstract

The TTG1-regulatory gene network regulates the development of all epidermal cell fates in *Arabidopsis thaliana.* Four members of the TTG1 complex, the serially duplicated R2R3-MYB *PRODUCTION OF ANTHOCYANIN PIGMENT* (*PAP*) genes, have previously been implicated in regulating the late stages of anthocyanin bionsynthesis in *Arabidopsis thaliana*. To study the effects of gene duplication, we sought to determine the extent of variation in each *PAP* gene compared to a single copy gene of the TTG1 network, *WEREWOLF*, using 48 naturally occurring *A. thaliana* accessions. It appears that the predominantly expressed *PAP1* gene demonstrates a biallelic pattern, consistent with other *A. thaliana* genes. All four genes fall below the average nucleotide diversity levels observed across *A. thaliana;* however, *WEREWOLF* demonstrates almost complete sequence conservation across the 48 accessions used in this study. We attempted to determine the relative ages of the four *PAP* genes, though this does not appear to correlate with accumulation of genetic variation. To investigate the genetic architecture of anthocyanin accumulation in *A. thaliana,* we performed an heritability and quantitative trait loci mapping analysis using a recombinant inbred line population derived from 19 natural *A. thaliana* accessions. While QTL were mapped for anthocyanin accumulation near several of the *PAP* genes, we observed a number of loci with no obvious candidate genes, providing novel insights into the genetic architecture of anthocyanin accumulation in *A. thaliana.* This work contributes to a greater understanding of the roles of regulatory genes in biosynthesis and the molecular basis of regulation as well as the effects of gene duplication on nucleotide variation in the resulting genes.

2. Introduction

Anthocyanins are a physiologically important group of plant pigments, the product of a specialised branch of the flavonoid biosynthetic pathway. Anthocyanins have been subject to intensive research in the past, which continues today, due to the wide array of conditions under which anthocyanin accumulation occurs and the present range of biological roles which they are involved in. As a result, the chemistry and synthesis of anthocyanin production is well characterized, though the roles and mechanisms of regulation of anthocyanin accumulation remains the subject of much research.

As pigments, anthocyanins confer a range of colours from red through to dark blue determined by a range of factors such as the chemical structure of the molecule, abiotic influences and conditions within the cell (Mathur *et al*., 2010). It is estimated that there are more than 400 naturally occurring anthocyanin species which vary around a common molecular structure. These nuances differentiate them into a number of semi-redundant roles (Kong *et al.*, 2003). The most commonly occurring anthocyanin species in fruits and vegetables are perlargonidin, cyanidin, delphinidin, petunidin, peonidin and malvidin; the richest sources of these pigments for human consumption are red fruits, red wine, cereals and certain vegetables such as red cabbage (de Pascual-Teresa *et al.*, 2010). Anthocyanins have been implicated in a number of roles within a plant, from attraction of pollinators and seed distributors to photoprotection of foliage. Recently, with the advent of new and more efficient molecular analysis techniques, anthocyanins have come under renewed scrutiny due to their ubiquitous nature and the relative ease with which they can be studied *in vivo*. This has increased the scope of understanding in anthocyanin biosynthesis and regulation of induction and accumulation. Anthocyanins have acted as a case study in transcriptional regulation, as complex transcriptional regulatory elements act to regulate anthocyanin accumulation. Further, epidermal cell traits themselves make for convenient phenotypic studies as effects of transcriptional regulation on these traits quite literally come to the surface.

Anthocyanins are phenolic compounds belonging to the flavonoid group which display a range of pigmentations from red to blue (the Greek *anthos-kyanos* meaning "blue flower") (Kong *et al*., 2003; Mathur *et al*., 2010). Structurally speaking, an anthocyanin molecule is a colourless anthocyanidine moiety with an attached glycoside group (Mathur *et al*., 2010). The observed pigmentation produced by the anthocyanin molecule is the work of the glucosidic group interacting with radiant light; the colour unique to an anthocyanin species is dependent on the pattern of oxidation, glycosylation and dehydration and cellular conditions (de Pascual-Teresa *et al*., 2010; Mathur *et al*., 2010). Anthocyanins are stable under mildly acidic (pH ≤5.0) conditions but show degradation and reduced biological efficacy under sustained alkaline (pH ≥7.0), high temperature (≥80°C), low temperature (≤0°C) and high light/UV radiation environments (Vollmannova *et al*., 2009; Wang *et al.*, 2010).

2.1 The Biological Roles of Anthocyanins

Anthocyanin accumulation in flowers, seeds and fruit has been widely accepted as a mechanism for attracting pollinators and seed dispersers (Holton & Cornish, 1995). Anthocyanin accumulation is also regarded as a generic stress response common in flora, as a number of phenomena induce anthocyanin accumulation. Such phenomena include nutritional stress, senescence of deciduous leaves during autumn, leaf expansion, ultraviolet light stress, radiant light stress and response to herbivory and insect or fungal attack (Atkinson, 1973; Longsreth & Nobel, 1980; Bongue-Bartelsman & Phillips, 1995; Thiele *et al.*, 1998; Hoch *et al.*, 2001; Vanderauwera *et al*., 2005; Feyissa *et al.*, 2009). While accumulation under these conditions is well documented and easily observed, the reason why anthocyanin accumulation occurs under these conditions is often contentious. Despite this, anthocyanin accumulation is predictable and consistent within and across species. Interestingly, accumulation of anthocyanin pigments as a general response across a number of taxa and conditions led Matile (2000) to propose anthocyanin accumulation serves no immediate function, but rather demonstrates "a kind of extravagancy without a vital function," rather than a specific response to the prevailing conditions.

Still, a number of conditions under which anthocyanin accumulation can be predicted to occur have been studied and documented. Nutrient stress is often cited as a key cause of anthocyanin induction and accumulation. Association of anthocyanin accumulation with plants experiencing nutrient deficiency in NO_3^-, PO_4^{2-}, or K^+ ions, three basic ions necessary for photosynthetic homeostasis, has been previously demonstrated (Longsreth & Nobel, 1980). Bongue-Bartelsman and Phillips (1995) demonstrated that nitrogen deficient conditions cause an up-regulation of genes involved in flavonoid and anthocyanin production. Feyissa *et al*. (2009) also noted an up-regulation of the bHLH gene *GLABRA3* (*GL3*) and the MYB genes *PRODUCTION OF ANTHOCYANIN PIGMENTS*(*PAP*)*1/2* in response to nutrient stress and an associated increase in anthocyanin accumulation. Similarly, phosphate deficient conditions resulted in increased anthocyanin accumulation across a number of taxa (Atkinson, 1973).

Light attenuation has also been cited as a common reason for anthocyanin accumulation in plants (Steyn *et al.*, 2002). Anthocyanins naturally occur in small amounts, evenly distributed throughout leaf tissue. They absorb light within the photosynthetically active spectrum between 400nm and 600nm (Close & Beadle, 2003). Anthocyanins also demonstrate transient accumulation coincidental with light exposure, with accumulation directed to the leaf periphery, areas of high light incidence (Steyn *et al*., 2002). It is proposed that they are involved in preventing over-excitation of photosynthetic pigments by absorbing excess radiation that would otherwise be absorbed by chlorophyll (Tucic *et al.*, 2009).

A familiar case of anthocyanin accumulation is senescent leaves during autumn. The proposed explanation of such phenomena is photoprotection of photosynthetic pigments (Hoch *et al.*, 2001), as over-stimulation of photosynthetic pigments that are being reabsorbed into the plant before abscission can result in photoinhibition (Powles, 1984). Anthocyanins reflect photosynthetically-active red light off the leaf, likely giving the photosynthetic pigments some reprieve in their semi-dormant state. Similarly, anthocyanin accumulation is documented to occur in conjunction with the

expansion stage of developing leaves (Kursar & Coley, 1992; Drumm-Herrel & Mohr, 2006). Foliar photostability is achieved once photosynthetic pigments are able to consistently resist light-induced degradation; this state is only achieved once a specific amount of chlorophyll is accumulated and may occur gradually with leaf expansion or may occur once the leaf has fully expanded, depending on both the species and environmental conditions (Kursar & Coley, 1992; Drumm-Herrel & Mohr, 2006). Drumm-Herrel and Mohr (2006) propose anthocyanins stave off excessive photosynthetically active radiation until the leaf reaches photostability where photosynthetic pigments are able to efficiently process incident light. Further, it was demonstrated that accumulation of anthocyanins in expanding leaves gives greater photostability in the juvenile stages (Drumm-Herrel & Mohr, 2006). Anthocyanins, though, do not act alone; they work in harmony with other flavonoids to provide a unified photoprotective function (Gould *et al.*, 2000; Steyn *et al.*, 2002).

High light levels are considered ideal for fruit growth and development as it results in an increase of phenolic content, especially anthocyanins (Jakopic *et al.*, 2009). Whether radiant light has a direct or indirect influence on anthocyanin accumulation is unknown, though there is supporting evidence for both theories. High light levels correlate with an elevation in free radicals and oxidative species levels as a natural by-product of increased photosynthetic activity. This increase in reactive oxygen species is known to up-regulate genes involved in oxidation defence, including anthocyanin biosynthetic genes, to prevent cellular damage (Vanderauwera *et al*., 2005). On the other hand, expression profiles show an increased concentration of proteins involved in anthocyanin production and accumulation not synchronised with other cellular responses to high light conditions, indicating anthocyanin accumulation may be regulated as a direct result of high light exposure, rather than as an indirect response co-ordinated with other light responses (Vanderauwera *et al*., 2005).

In either case, there is an ever increasing body of evidence supporting the antioxidant activity of anthocyanins. Wang *et al*. (1997) tested five naturally abundant anthocyanin forms to demonstrate their effective antioxidant activity. Anthocyanins were found to be at least as effective in their

oxygen radical absorbance capacity, with the most effective being three and a half times greater than Trolox, a synthetic commercial antioxidant. Gould *et al.* (2000) demonstrate the predominance of anthocyanin accumulation in the leaf mesophyll and propose that this makes them ideally situated to scavenge oxygen radicals resultant of chloroplast activity. Vuleta *et al.* (2010) demonstrate seasonal fluctuations in anthocyanin accumulation, where anthocyanin production is greater during summer compared with spring or autumn, supporting the role of anthocyanins in free radical scavenging, as anthocyanin accumulation is greatest when photosynthetic activity, and as a result, reactive oxygen species production, is highest. Gould *et al.* (2000) demonstrate the ability of anthocyanins to mitigate rapid and dramatic increases in cellular H_2O_2 levels resulting from mechanical leaf injury. Margins around the site of damage high in anthocyanins were able to rapidly reduce and maintain low H_2O_2 levels, whereas regions bereft of anthocyanins accumulated oxygen radicals for several minutes after chlorophyll rupture. Similar results to those obtained from damaged leaves high in anthocyanins were yielded when cells lacking anthocyanins were treated with a commercial antioxidant after leaf damage (Gould *et al.*, 2002). Sarma and Sharma (1999) observed anthocyanins acting to protect DNA in vivo, where an interaction between bovine DNA and anthocyanins reduced degradation of both components involved when subjected to free radicals, whereas free radical exposure resulted in severe oxidative damage to each component separately.

A link between high UV radiation levels and anthocyanin accumulation has been proposed, though this has been argued as correlative rather than causal, since anthocyanins have minimal UV-A and UV-B absorbance capacity (Lindoo & Caldwell, 1978; Beggs & Wellmann, 2008). Further, Stintzing and Carle (2004) note that absorption of radiation with wavelengths between 280-320nm is primarily by hydroxyl-cinnamoylated molecules, while a number of anthocyanin groups are glycosylated. However, there are a number of anthocyanin groups acylated with cinnamic acids, potentially placing them in range of UV absorbance capacity, and Burger & Edwards (1996) demonstrate that under high UV conditions, plants with an increased anthocyanin content maintain a higher photosynthetic rate and increased net quantum yield compared with plants of a lower

anthocyanin content. Shade plants, which have a naturally low anthocyanin content compared to plants in high light environments, demonstrate increased photosynthetic efficiency when UV-A and UV-B light is filtered out during full light exposure; further, shade plants which have anthocyanins artificially induced via UV-B exposure demonstrate reduced photoinhibition compared to plants without increased anthocyanin content (Thiele *et al.*, 1998). It therefore seems likely that while anthocyanins do not have considerable native UV absorbing capabilities, they contribute to overall photostability under UV stress and reduce photoinhibition.

Anthocyanin accumulation has been suggested as a possible response to herbivory and insect attack, though some have concluded that this too is correlative rather than causal. Costa-Arbulu *et al*. (2001) demonstrated that although leaves under attack by aphids showed increased anthocyanin accumulation at the site of the attack, water-stress induced a similar response. Interestingly, aphid fecundity was reduced after being subjected to a diet of post-attack anthocyanin-rich leaves, though an artificial anthocyanin-rich diet given to the aphids outside of a leaf medium did not produce the same result (Costa-Arbulu *et al*., 2001). Moreno *et al.* (2010) demonstrate the plant signalling hormone methyl jasmonate, produced in response to both biotic and abiotic stresses, causes an increase in anthocyanin accumulation. The mechanism through which anthocyanins act to protect plants in these cases, if one exists at all, are yet to be elucidated. Anthocyanin accumulation also occurs in response to fungal attack, but only once pathogen growth has been inhibited. Based on this, Close and Beadle (2003) suggest fungal resistance is conferred by unrelated physiological or chemical processes, while anthocyanins act in a photoprotective role for tissue damaged in the attack rather than as an anti-fungal agent.

A controversial case for cellular osmotic adjustment by anthocyanin accumulation is presented by Close and Beadle (2003), where they highlight that although anthocyanin accumulation is observed in drought and low temperature conditions, the studies that support this are not conclusive as to whether anthocyanin accumulation is involved in osmotic adjustment or photoprotective roles. It is

also noteworthy that considerable anthocyanin damage and degradation occurs under freezing conditions (Vollmannova *et al.*, 2009) so anthocyanin production and accumulation would be a constant and energy inefficient response to such stresses. Lovisolo *et al*. (2010) also note that drought stress results in anthocyanin accumulation, though they too question whether this is a mechanism for osmotic adjustment or photoprotection by light attenuation. Interestingly, abscisic acid, the plant hormone involved in the opening and closing of stomata, is also directly involved in the biosynthetic pathway of anthocyanins, providing a physiological link between drought-tolerance and anthocyanin accumulation (Lacampagne *et al.*, 2010). While there is an apparent link between stress conditions requiring stomatal aperture regulation and anthocyanin production, whether the accumulation of anthocyanins acts in an osmotic adjustment or light attenuation role is yet to be determined.

2.2 Anthocyanin Biosynthesis and Regulation of Production

The flavonoid biosynthetic pathway has been well characterised and is highly conserved amongst plant species (Springbob *et al.*, 2002). The pathway is divided into two sections: the 'early stage' genes, from *C4H* to *CHI*, wherein the basic structure of all flavonoids is produced; the 'late stage' genes, from *F3H* to *GT*/*ACT*/*MAT*, wherein basic flavonoid structure is modified to produce specific anthocyanin types (Martin *et al.*, 1991). The 'late stage' genes are the genes targeted for regulation by the TTG1/bHLH/PAP regulatory protein complex in *A. thaliana* (Gonzalez *et al.*, 2008) (*Figure 1*). This early/late stage division allows efficient regulation of production as the generic flavonoid precursor can be produced independent of the flavonoid modification processes. It has been proposed that the structural proteins of the biosynthetic pathway form protein complexes and compete for substrates, allowing a rapid reaction to substrate availability (Springbob *et al.*, 2002). Formation of protein complexes confers stereospecificity; association between the proteins involved in forming the basic flavonoid structure and those involved in the modifying stages have been demonstrated, indicating that selective association between structural proteins pushes the

biosynthetic reaction and use of available substrate in the favour of a desired pigment structure (Springbob *et al.*, 2002).

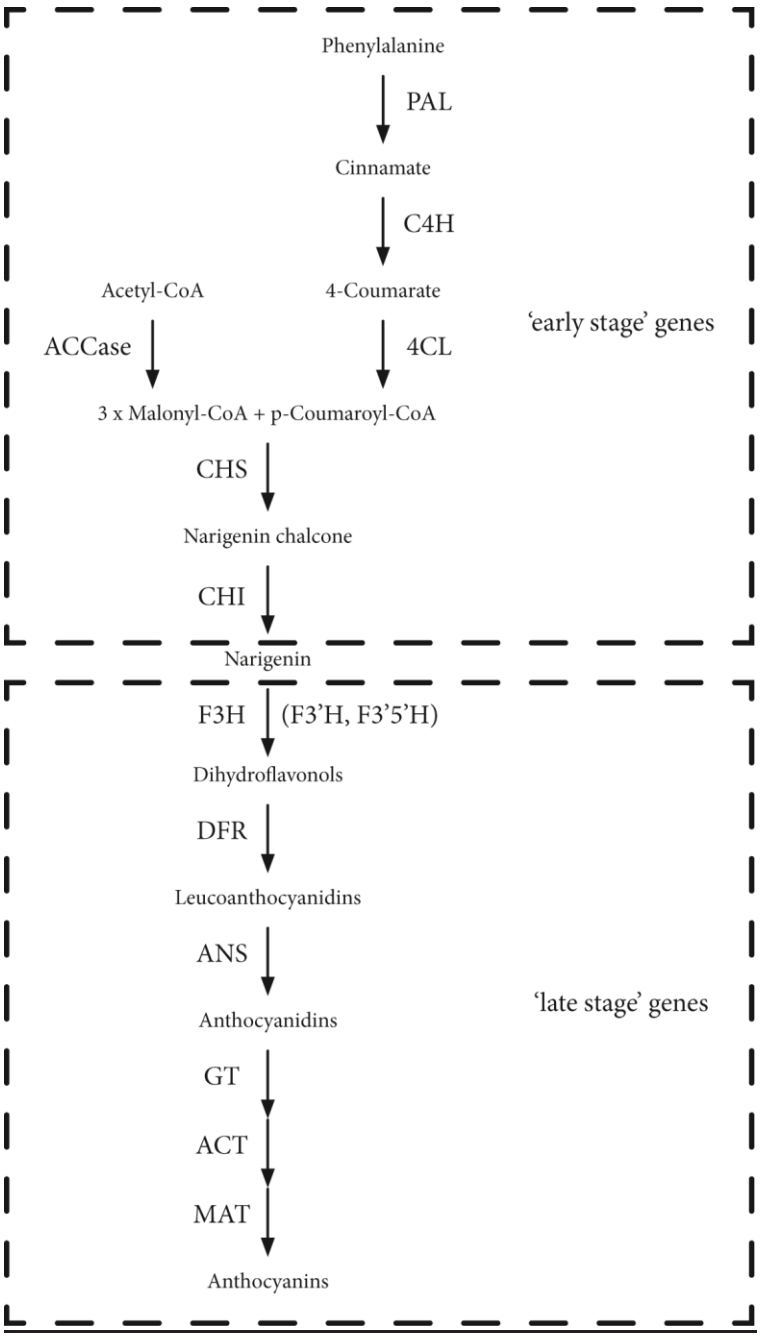


Figure 1 Scheme of the anthocyanin biosynthetic pathway. ACCase, acetyl-CoA carboxylase; PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate:CoA ligase; CHS, chalcone synthase; CHKR, chalcone ketide reductase; CHI, chalcone isomerase; F3H, flavanone 3β-hydroxylase; F3'H, flavonoid 3'-hydroxylase; F3'5'H, flavonoid 3',5'-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; GT, glucosyltransferase; ACT, anthocyanin acyltransferase; MAT, malonyltransferase. Figure modified from Springbob *et al.* (2003).

Anthocyanins themselves are produced via a specialised branch of the flavonoid production network, all of which are produced using the same two precursors, malonyl-CoA and p-coumaroyl-CoA. Anthocyanidin synthase oxidises leucoanthocyanidins to form anthocyanidins, the un-glycosylated precursor of all anthocyanins. Anthocyanidins are modified to produce specific anthocyanins by glycosylation, oxidation and dehydration, which is regulated by three enzymes at the end of the anthocyanin production pathway, glucosyltransferase**,** anthocyanin acyltransferase and malonyltransferase, respectively (Lacampagne *et al*., 2010).

It has been demonstrated that sucrose has the capacity to induce anthocyanin accumulation. Interestingly, it appears that anthocyanin accumulation is a direct response to sucrose exposure, rather than part of a stress response induced by sucrose exposure; treatment of *Arabidopsis thaliana* with mannitol activated *KIN1*, a stress-induced gene, whereas sucrose treatment did not (Solfanelli *et al.*, 2006). It is proposed that sucrose itself acts as a signal for anthocyanin accumulation, as accumulation of sugar in *Arabidopsis thaliana* leaves showed a 10- to 20-fold increase in expression of structural genes of the anthocyanin biosynthesis pathway. *PAP1* is the only regulatory R2R3 MYB gene shown to be up-regulated in response to sucrose treatment, and genes affected by sucrose treatment were unresponsive to glucose and fructose treatments. Sucrose acting as an anthocyanin accumulation signal appears to be a sound proposal, as concentrations as low as 7.5mM are enough to elicit an increase in cellular mRNA levels of the genes involved in anthocyanin biosynthesis (Solfanelli *et al*., 2006).

Increasing the complexity of a case study in anthocyanin production, there is also a range of factors which have an indirect effect on anthocyanin biosynthesis and accumulation. For example, the *Phalaenopsis* locus *CYP78A2* up-regulates the already active anthocyanin biosynthetic pathway in transformed petals, though it is not expressed in petals under normal developmental conditions (Su & Hsu, 2010). It appears that this up-regulation is an indirect effect, as anthocyanin accumulation occurs in petals naturally in absence of *CYP78A2* expression. It is therefore suggested

that the regulation of abscisic acid (ABA) and other plant hormones by *CYP78A2* is the mechanism responsible for the observed up-regulation of anthocyanin accumulation (Su & Hsu, 2010). Backing up this claim is the finding that exogenous applications of ABA up-regulate *MYB, PAL, CHI* and *CHS* gene expression in *Vitis vinifera* (Lacampagne *et al*., 2010); methyl jasmonate/jasmonic acid has been shown to up-regulate expression of genes involved in anthocyanin biosynthesis in strawberries (Moreno *et al*., 2010); salicyclic acid has been shown to increase anthocyanin accumulation (Mihai *et al.*, 2010); and a combination of abscisic acid and salicyclic acid in a two-stage media system has been shown to enhance anthocyanin production in *Vitis vinifera* (Mihai *et al.*, 2010).

Transcriptional regulation is the major controlling force behind anthocyanin biosynthesis (Yuan *et al.*, 2009). A wide range of factors influence anthocyanin accumulation, in reflection of the various roles they fulfil and the conditions under which their synthesis is induced. The transcription factors primarily involved in regulation of anthocyanin biosynthesis have been isolated in a number of species. bHLH (basic Helix-Loop-Helix) and R2R3 MYB proteins form a regulatory protein complex upon a WD-40 repeat residue scaffold (Zhang *et al.*, 2003). Across species, the proteins appear to have a similar molecular structure though the way they interact with their target genes differs (Yuan *et al.*, 2009).

Albert *et al*. (2010) demonstrated that anthocyanin production could be restored by introduction of combinations of bHLH/MYB gene constructs in a commercial orchid cultivar where lack of these transcription factors confers white flowers. Known bHLH/R2R3 MYB anthocyanin activators were taken from a number of species and introduced to the white orchid cultivar. When both bHLH and R2R3 MYB proteins were introduced together, anthocyanin production was restored. MYB proteins are directly involved in binding with the target DNA. Due to this, activation of anthocyanin production in the monocotyledon *Cymbidium* genus was much more effective when the MYB protein also originated from a monocotyledon, as the monocotyledon MYB proteins differ by a single residue from dicotyledons, affecting their ability to bind to the genes which they activate when they

are out of their specific context (Albert *et al*., 2010). Further, the work of Niu *et al*. (2010) in Chinese Bayberry (*Myrica rubra*) demonstrated that up-regulation of the MYB gene *MrMYB1* increases anthocyanin accumulation. However, the nonsense mutation *MrMYB1d* was unable to activate the anthocyanin biosynthesis pathway; this mutation is found in white fruit whereas the functional *MrMYB1* is expressed in red fruit.

In *Arabidopsis thaliana,* determination of epidermal cell fate is dependent upon three key protein families which form a regulatory complex: a WD40 repeat protein *TRANSPARENT TESTA GLABRA1* (*TTG1*), bHLH proteins and MYB proteins (*Figure 2*) (F. Zhang *et al*., 2003; Ramsay & Glover, 2005). TTG1 is the scaffold upon which the regulatory complex is built; TTG1 associates with a bHLH protein which associates with a MYB protein, though TTG1 and the MYB proteins do not physically interact. In the case of R2R3 MYB proteins, the R2 region of the MYB protein is involved in binding to the target DNA and the R3 region is primarily responsible for bHLH protein-protein association (Oda *et al*., 1997; Zimmermann *et al*, 2004). In the case of the anthocyanin biosynthetic pathway, three bHLH proteins have the capacity to regulate biosynthesis: GLABRA3(GL3), ENHANCER OF GLABRA3(EGL3) or TRANSPARENT TESTA8(TT8). While the four R2R3 MYB genes *PAP1*, *PAP2, AtMYB113* (from here referred to as *PAP3*) and *AtMYB114* (from here referred to as *PAP4*) have the potential to functionally complete the regulatory complex, the PAP1 protein predominantly interacts with the bHLH protein to complete the regulatory complex (*Table 1*) (F. Zhang *et al*., 2003). The necessity of these proteins for regulation of epidermal cell traits is demonstrated in mutant knock-out lines: *ttg1* mutants affect root hairs, trichomes, seed coat pigmentation, seed coat mucilage and anthocyanins. Similarly, *gl3*/*egl3*/*tt8* triple mutants are unable to express epidermal cell traits (F. Zhang *et al.*, 2003). With the advent of whole genome sequencing, it has become practical and highly informative to study, not just gene expression and function, but also natural variation in genotypes between individuals and populations. It has become apparent that genes and DNA regions affecting gene expression are highly variable and variation in gene expression over differences in protein forms has become a greater focus of study in evolution (Oleksiak *et al.*, 2002).

As such, the extent of this variation is being investigated in a number of taxa (de Bono & Bargmann, 1998; Cubas *et al.*, 1999; Johanson *et al.*, 2000; Kliebenstein *et al*., 2001; Oleksiak *et al.*, 2002; Cheung *et al*., 2003; Xue *et al*., 2008).

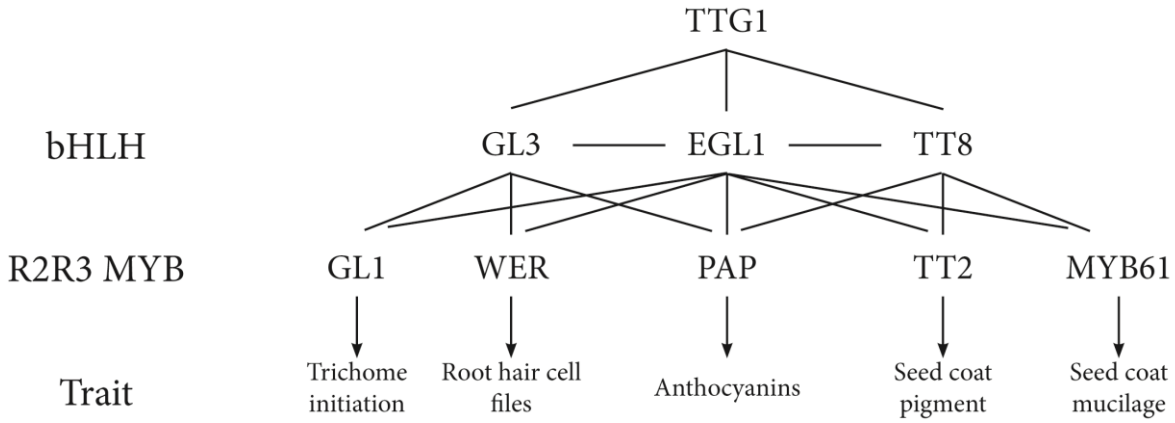


Figure 2 TTG1 regulatory network model. This modified figure (F. Zhang *et al.,* 2003) shows interactions between all known bHLH and MYB transcriptional regulators which determine epidermal cell fates in *A. thaliana*. Black lines signify demonstrated interactions between the proteins and genes. Arrows indicate the epidermal cell fate which the R2R3 MYB protein specifies.

Table 1 Gene Nomenclature

Gene Number	Gene Name	Common Name (used in this work)
At1g56650	*AtMYB75*	*PAP1*
At1g66390	*AtMYB90*	*PAP2*
At1g66370	*AtMYB113*	*PAP3*
At1g66380	*AtMYB114*	*PAP4*
At5g14750	*AtMYB66*	*WER*

In our work, we seek to contribute to the growing body of literature offering an ever-increasing understanding of how transcriptional regulation controls gene expression and the extent to which genetic variation impacts on this regulation. We use the four duplicate *PAP* transcription factor genes involved in regulation of anthocyanin biosynthesis as a case study in the fate of duplicated genes. *A. thaliana* is a weedy species distributed across the world in a variety of habitats. It has been suggested that radiation of accessions from parent populations in Asia and the Iberian peninsula has occurred relatively recently (Innan *et al.*, 1996; Sharbel *et al.*, 2000). Because of its wide natural distribution, *A. thaliana* makes an ideal candidate species for a study in natural variation. We use a

collection of natural *A. thaliana* accessions from a range of environments known as the Versailles core-collection (McKhann *et al*., 2003) in an attempt to broadly sample the natural variation across the *A. thaliana* range and determine the extent of genetic variation in the four *PAP* transcription factor genes (F. Zhang *et al.*, 2003). The Versailles core-collection of accessions has been shown to sample the major genotypic populations of *A. thaliana* and is therefore a reliable tool for considering the genetic variation of the species (Simon *et al*., 2012). We also use quantitative trait mapping to identify the loci involved in the variation of anthocyanin accumulation, both to confirm the involvement of the *PAP* genes as well as investigate other genes that may not yet have an involvement in anthocyanin accumulation ascribed to them. For this, we use a Multiparent Advanced Generation Inter-Cross (MAGIC) population, again to capture as much variation as possible in a single inbred population (Kover *et al*., 2009). Our work has the potential to elucidate the functional capacity of the *PAP* genes other than *PAP1* and whether they likely continue to play a role in regulation of anthocyanin accumulation as we identify the extent of variation across the duplicated *PAP* genes. Further, we will be able to provide some insight into whether the observed natural variation in anthocyanin accumulation is predominantly the work of the *TTG1* network, or whether there are other factors involved. Overall, our work will contribute to a further understanding of regulation of anthocyanin biosynthesis and the resulting natural variation of anthocyanin accumulation, as well as an understanding of the fate of duplicate genes and the genetic basis of natural variation in general.

3. Molecular Analysis of the *PAP* Genes

3.1 Introduction

Gene duplication has long been held as an important source of material for biological evolution to act upon (Ohno, 1970; J. Zhang, 2003). Over the past 20 years, one source of mass gene duplication, polyploidy, has been extensively studied and well established as a common event in plants, with up to 80% of angiosperms being polyploid (Leitch & Bennett, 1997; Irish & Litt, 2005). A less commonly studied source of genetic material for selection to act upon, segmental duplication (Simillion *et al.*, & Van de Peer, 2002; Cannon *et al.*, & May, 2004; Leister, 2004), has been shown to be particularly important in the evolution of large gene families, such as the *TTG1* pathway (Zhang *et al.*, 2003) and the MYB transcription factor superfamily (Yanhui *et al*., 2006). Since the complete sequencing of the *Arabidopsis thaliana* genome (The Arabidopsis Genome Initiative, 2000), it has been observed that the genome is replete with paralogous genes, the proposed origin of this being at least a single whole-genome duplication as well as large-scale segmental duplications (Simillion *et al*., 2002; Raes *et al.*, 2003). No doubt due to these duplication events, roughly 5% of the *A. thaliana* genome has been proposed as being dedicated to transcriptional regulator genes (Riechmann *et al*., 2000); the complex *A. thaliana* MYB superfamily of transcriptional regulators involved in regulation of development is thought to have been a result of rapid expansion in the past (Yanhui *et al*., 2006), and for this reason has been, and still is, the subject of much investigation into the evolution of large complex gene families (Kranz *et al*., 1998; Riechmann *et al*., 2000; F. Zhang *et al*., 2003; Zimmermann *et al*., 2004; Dubos *et al*., 2010). For the most part, previous studies of the *A. thaliana* MYB family have taken the approach of knocking out and/or overexpressing genes to ascribe function to them (Gonzalez *et al.*, 2008). However, the previously established high incidence of genetic duplication combined with the wide geographic distribution of *A. thaliana* (Scmuths *et al.*, 2004) means phenotypes vary greatly throughout the natural *A. thaliana* population; genetic knock-out or overexpression analyses may not capture the entire scope of genetic effects when this

potential for high genetic variation is taken into account (Alonso-Blanco *et al*, 1998; Bergelson *et al*, 1998; Li *et al.*, 1998; Koorneef *et al.*, 2004).

Here, we examine natural variation in the four duplicated R2R3 MYB transcription factor genes *Production of Anthocyanin Pigment* (*PAP*)*1, PAP2, PAP3* and *PAP4,* originally identified via reverse genetics studies, functional characterisations and motif searches (Kranz *et al*., 1998; Borevitz *et al.*, 2000; Stracke *et al.*, 2001). All four *PAP* genes are all located on chromosome 1. *PAP1* is located 3.52 mbp from, and is in the reverse orientation to, the other *PAP* genes; *PAP3, PAP4* and *PAP2* (named in the order they occur on the chromosome) are located consecutively within 10 kbp of each other. For comparative purposes, we also examined the related single-copy R2R3 MYB gene *WEREWOLF* (*WER*), which acts in the same epidermal cell fate pathway. R2R3 MYBs rely on a particular structure in order to accurately interact with bHLH proteins and their target DNA. The typical structure is two imperfect amino acid sequence repeats, R2 = 50 amino acids long, R3 = 50 amino acids long. Each repeat forms thee α-helices, with the second and third helices of each repeat forming a helix-turn-helix structure. The third helix is the recognition helix which makes contact with the target DNA (Stracke *et al*., 2001; Dubos *et al*., 2010). In the *A. thaliana* MYB family, the repeats are evenly distributed with three tryptophan residues each; however, in the case of the *PAP* genes, the first tryptophan residue for the R3 repeat sequence is consistently replaced with a phenylalanine residue (Dubos *et al.*, 2010). The *PAP* gene family is involved in initiating anthocyanin accumulation, while *WER* is involved in regulating root hair initiation. Both the *PAP* genes and *WER* are members of the *A. thaliana TTG1* epidermal cell fate pathway; the effects of genetic variation are readily discernible in their phenotypes, making them ideal candidates for a study in genetic variation of a complex of genes.

Initiation of the anthocyanin production pathway requires the PAP proteins to associate with the functionally overlapping bHLH transcription factors GL3, EGL3 and TT8 and the eponymous WD-40 repeat factor TTG1 to form an activation complex; ectopic *PAP* expression does not result in

anthocyanin accumulation without the presence of bHLH/TTG1 proteins (Shi & Xie, 2010). Loss of function of either *TTG1* or *PAP1* will result in reduction in anthocyanin accumulation, though *ttg1* mutants affect a host of epidermal phenotypes (Koorneef, 1981), while *pap* mutants are specific to anthocyanin accumulation (Gonzalez *et al*., 2008). While *PAP1* has been shown to be predominantly expressed in both seedlings and mature rosettes (Gonzalez *et al.*, 2008; Shi & Xie, 2010) as well as being up-regulated under conditions inducing anthocyanin accumulation (Teng *et al.,* 2005; Lea *et al.*, 2007; Lillo *et al.*, 2008; Rowan *et al*., 2009), *PAP2*, *PAP3* and *PAP4* expression is negligible by comparison. Still, *PAP2*, *PAP3* and *PAP4* do have the capacity to recover *pap1* mutants, and complete knockdown of expression of the four *PAP* genes prevents anthocyanin accumulation completely (Gonzalez *et al*., 2008), demonstrating some degree of functional redundancy. Further, *PAP2* expression has been shown to increase under certain environmental stress conditions (Mission *et al*., 2005; Morcuende *et al*., 2007; Muller *et al.*, 2007; Lillo *et al*., 2008) and anthocyanin accumulation in seed coat testa appears to occur independent of *PAP1* expression (Gonzalez *et al*., 2008), suggesting potential variation in timing, location and circumstance of expression of the *PAP* genes.

Anthocyanins are a specialised group of flavonoids, products of the flavonoid biosynthetic pathway. They differ due to unique changes conferred by enzymes at the end of the flavonoid biosynthetic pathway (Lacampagne *et al.*, 2010). The wide array of conditions under which anthocyanin accumulation is induced and the range of biological roles in which they are involved has resulted in them being subject to intensive research. This has led to a comprehensive understanding of their chemistry, synthesis, regulation and roles within the plant. Anthocyanins confer a range of colours from red through to dark blue, determined by a range of factors such as the chemical structure of the molecule, abiotic influences and cellular conditions (Mathur *et al*., 2010). The current model of anthocyanin production proposes the involvement of 11 protein families: PAL, C4H, 4CL, CHS, CHI, F3H, DFR, ANS(/LDOX), GT, ACT and MAT (listed from the start of the pathway to the end) (Springbob *et al*., 2002; Shi & Xie, 2010). It appears that the proteins of the biosynthetic pathway targeted by the *PAP* genes are the 'late' stage proteins, starting from F3H down; these

proteins are down-regulated in the absence of the *PAP*-bHLH-WD-40 activation complex (Gonzalez *et al*., 2008; Shi & Xie, 2010). Natural populations and individual accessions of *A. thaliana* demonstrate both qualitative and quantitative variation for anthocyanin accumulation in leaves, siliques, stems and seeds but the molecular genetic basis of this variation remains unclear (Hered, 1989; Deikman & Hammer, 1995; Focks *et al.*, 1999; Cominelli *et al*., 2008).

Using a broad sample of *A. thaliana* accessions to examine natural variation in the *PAP* genes and compare and contrast to *WER*, we are able to examine: (i) how much variation is found in the sequence of these duplicated genes, and how that compares to a single copy gene in the same gene complex? (ii) Does variation in the *PAP* genes appear neutral, indicating the genes are functionally redundant and therefore no longer under selective constraint, or are patterns of mutation apparent, indicating selective pressures still acting on the genes? (iii) Has the relative age of the *PAP* genes affected the rate at which they have accumulated mutations? Our results better define the roles of the *PAP* genes, expand on our understanding of their evolutionary history and contribute to an understanding of the effects and results of gene duplication and its potential to act as the 'raw material' upon which biological evolution is able to act.

3.2 Materials and Methods

3.2.1 Plant Materials. A nested core collection of 48 natural *Arabidopsis thaliana* accessions was used to include maximum possible genetic diversity with minimal repetition (McKhann *et al*., 2003). Natural accessions of *A. thaliana* indicate seed collected from a single natural population (Appendix 1). Given the self-pollinating nature of *A. thaliana,* individuals from a single accession typically represent a single genotype. The genome of Columbia (Col-0) is not included in this core collection but was used for comparative analysis. Col-0 sequences were acquired from The *Arabidopsis* Information Resource (TAIR) (Lamesch *et al*., 2010). Approximately 5 seeds for each accession from the Versailles Core-collection were planted in 4-cm cells of seed raising mix (Oderings Nuseries, NZ) in 72-cell flats (Hummert International, Inc.). The accessions were sprayed with 1.8g/L (w/v)

Terrachlor fungicide solution and stratified at 4°C for five days. Flats were then moved to 25°C under 15-hr light and thinned to one plant per cell 7 days post-germination. Leaf tissue was sampled for DNA extraction six weeks post-germination.

DNA was extracted using a modified CTAB extraction protocol (Doyle & Doyle, 1987). Two or three fresh young leaves were taken from each accession and shaken in a beadmill at 5000rpm for 90 seconds in tubes containing 6 zirconium beads. 450µL pre-warmed CTAB was added to each tube, which were then shaken to mix. The samples were incubated at 65°C in a water bath overnight. After incubation, 800µL chloroform/isoamyl alcohol (24:1 v/v) was added and samples were shaken vigorously until emulsion formed. Samples were centrifuged for 5 minutes at 13.5krpm. The aqueous layer was transferred to new tubes and emulsified with 800µL chloroform/isoamyl alcohol (24:1 v/v), then centrifuged for 5 minutes at 13.5krpm. The aqueous layer was transferred to new tubes and 400µL of ice-cold isopropanol was added. The tubes were inverted 5 times to mix and placed in -20°C freezer for 30 minutes. The samples were centrifuged for 20 minutes at 13.5krpm. The supernatant was discarded and 500µL 70% Ethanol was added to the pellet. Samples were inverted 5 times to mix then centrifuged for 5 minutes at 13.5krpm. Supernatant was discarded and tubes were left open and inverted on paper towels overnight to air dry. 100µL filter-sterilized TE buffer was then added to each sample and shaken to re-suspend pellet.

3.2.2 PAP and WER sequencing. Primers for both polymerase chain reactions (PCR) and sequencing were designed manually or using PRIMER3 in GENEIOUS v5.3.4 (Drummond *et al*., 2010) from an alignment of the four PAP genes from the *A. thaliana* accession Col-0, acquired from TAIR (Rozen & Skaletsky, 2000). Approximately 500bp both up- and down-stream of each gene was included to allow production of flanking primers. A list of primers can be found in Table 2. PCRs were carried out using NEB *Taq* polymerase (New England Biolabs), Firepol DNA polymerase (Roche) and Acuprime DNA polymerase (ABI) according to standard PCR protocols (*Table 3*). All PCR reactions were visualised by gel electrophoresis using a 1% agarose gel run at either 75v for 60 minutes or

Table 2 List of primers used in PCR and sequencing reactions

PAP1	Primer sequence	*PAP2*	Primer sequence
P1 F1	CCTCCAATCATTTGTGAAACC	P2 F1	GTCATTTTCCTATGC
P1 F2	CAAACATGCACGTCACTTC	P2 F2	TGGATATCAAACATGCACGTC
P1 F3	GCACCAAGTTCCTGTAAGAGC	P2 F3	GCATCAAGTTCCTTTGAGAGC
P1 F4	TGAAAATTTCTTTTGCTGTTCG	P2 F4	TTATGTAGGGCTAAATCGATGC
P1 New F1	CCTTTTACAATTTGTTTAT	P2 New F1	AGTTTTAGAATAATCCGAT
P1 New F2	TATAAATTTTCTCACATAC	P2 New R1	ATTCTTTACTTCATACAGA
P1 New F3	ATATCAAACATGCACGTCAC	P2 New R2	TACGAAAGAAAAGCCACC
P1 New F3	ATATCAAACATGCACGTCAC	P2 R1	CCGTCTTTCATTGAATCTTGC
P1 New R1	TTCTTCAAAATGTAAACACGG	P2 R2	AAAGTTGCTCAACGTCAAACG
P1 New R2	TTTACTGTTGTCATGTAG	P2 R3	CCTTCCACCATGTCTACTTGCC
P1 R1	AACATGACATTTGCCCAACCA	P2 R4	AGGTCGAGGCTTAAAAACACC
P1 R1.n	GAAGCTAGCAATACGCAACG	P2F1.1	CATTATCAATAAGTCCTCCCAGG
P1 R2	GGTCGGTAGAGAAGGTTTGG	P2FFlank1	AACGGCCGGCTCAACGGG
P1 R3	ATCAACGTCAAAAGCCAAGG	P2FInt1	GCTGGTCGATTGCCTGGTCGG
P1 R4	CAGTAATTCTTGACGTCATTTGC	P2FInt2	CGACGACAGCTGAACATGGGGC
P1Fcheck1	CCTCGATACTACTTAACTCGGAGAG	P2R1.1	CACTCACCAGAAGACAGCCC
P1Fcheck2	CGACTGCAACCATCTCAATGCCCC	P2RFlank1	GTCTGTGTGTGTGTGGTGGGG
P1Fcheck3	CAGTACCAAACCTTCTCTACCGACC	P2RInt1	GGACCGGTGTTGTAGGAGGGG
P1Fcheck4	CGTGCAGATTCTCATATGGGC	P2RInt2	TCGTCGCTTCAGGAACAATCGC
P1Rcheck1	CCTAGAAGGTTGGGAAGGCG	P2RInt3	CGTCAAACGCCAAAGTGGCCC
P1Rcheck2	GGAAACTATCTCCATCACAAGGG	Seq-P2-F1	GGATTTGGGAAGCCACAATAACC
P1Rcheck3	CTAACTCCATCCGTAATTCTTACC	Seq-P2-R1	GACTTACAACACGAAGAC
P1Rcheck4	CCATGCATTTATAGTCCAACCGG		
Seq-P1-F1	GGGTTCGTCCAAAGGGCTGC		
Seq-P1-F2	GATAGTCTCTTGAGACAGTGC		
Seq-P1-R1	GTGATTAGAAATATAGGCGG		
Seq-P1-R2	CTGTTGTCATGTAGATAACC		
PAP3	Primer sequence	*PAP4*	Primer sequence
P3 F1	CTTTGCTCCATGGGCGAATCAC	114midR1	CAATGTACTATACACACATACG
P3 F2	GATATTCTCTTGAGGCAATGC	114midR2	GTTAAACCTAAGAAGATTGCTG
P3 F3	CGATTGGTGTGCATATATTC	P4 F1	CTCTACTCCTCCAAGGATCTC
P3 New F1	ATTTTATAAAATTTGTGAG	P4 F4	CTACGATTGGTTTGTCTTC
P3 New F2	CTCTTGTCTAAACGTGGAT	P4 R1	GTAAACTCACACAGAAACGG

PAP3 (*cont.*)	Primer sequence	*PAP4 (cont.)*	Primer sequence
P3 New F3	GAACGAAGAGTCTGGCTGC	P4 R4	CCTACGTATACATACGAG
P3 New R1	GTTATATTGTAACAAATGCTG	P4FFlank1	AGCCTGAATTTTGACCGTTATGTGGG
P3 New R2	AATATAAGTAGAAACGACTC	P4FFlank2	TTAGTCCCCCTTGAGGTTTCTGAGG
P3 R1	CTTCGTCTTACAGCATCGTTC	P4FFlank3	GGCTAGAGTGGGGAAGTTCAATTCTGC
P3 R2	CCACCTAAACATTGTGAAAC	P4FInt1	GGGGTTGGTCCGAAGTTGTTCCG
Seq-P3-F1	GGCTGCTAAGTGCAGTGAACC	P4FInt2	TGCTGGTCGATTACCTGGTCGG
Seq-P3-R1	CCATTAGACGTTATCTTCTCGGG	P4FInt3	CCCCTCCTAATACACCGGCCC
Seq-P3-R2	CATATTATCTCCTAATGAACC	P4RFlank1	TTCCGTTCTCCGGTTGGCCC
		P4RInt1	CAACTTTTCCTGCACCGATTTAGCCC
		P4RInt2	TGGCCCTCGTCTAGCAAACGC
		P4Start1	CCATGGAGGGTTCGTCCAAAGGG
		Seq-114-F1	GCCCGTCACGTGTGTATATATTTCC
		Seq-114-F2	CAGACTCGTTAGATTTTGATCCGG
		Seq-114-R1	CAGAAACGGAAACATAAGGTTTGCC
		Seq-114-R2	GGAGGGGTTATAATATTTATCC

150v for 25 minutes. Agarose gels were soaked in Ethidium Bromide before being photographed in a gel dock under UV light.

Prior to sequencing, PCR products were cleaned using a standard ExoSap clean up protocol. For each sample, 7.75 µL H_2O, 0.25 µL exonuclease 1, 0.5 µL shrimp acid phosphatase and 8 µL of the sample to be cleaned was mixed and heated at 37°C for 30 minutes, 80°C for 15 minutes then held at 4°C. PCR products were sequenced initially using a modified Sanger sequencing protocol (*Table 4*). On average, three µL of template was used for sequencing, though the template amount was adjusted to reflect the intensity of the band as visualised on a gel, with more being used for weaker gel bands less for stronger gel bands. After the sequencing reaction, samples were prepared for capillary separation using an ethanol clean-up protocol. Eight µL of Beckman Coulter Agencourt CleanSeq beads and 49.6 µL 85% Ethanol were added to the 16 µL Sequencing reaction product. The samples were pipetted up and down seven times to mix and were then left on a magnetic plate for four minutes. The supernatant was removed and 100 µL 85% Ethanol was added to wash. The samples were again left on a magnetic plate for four minutes. The supernatant was removed and the samples were left to air-dry for 30 to 40 minutes. 40 µL 0.1mM EDTA was added to elute and samples were pipetted up and down 7 times to mix. Samples were left for five minutes at room

temperature then placed back on magnetic plate for 4 minutes. 30 µL of the sample was put in a clean tube for submission for capillary separation and analysis by the Massey Genome Service, Turitea Campus, Palmerston North.

Table 3 Standard Polymerase Chain Reaction protocols

NEB taq recipe	Firepol taq recipe
15.3 µL H_2O	13.3 µL H_2O
2 µL Buffer	2 µL Buffer
0.5 µL dNTPs	2 µL $MgCl_2$
0.5 µL Forward primer (20 µM)	0.5 µL dNTPs
0.5 µL Reverse primer (20 µM)	0.5 µL Forward primer (20 µM)
0.2 µL NEB taq polymerase	0.5 µL Reverse primer (20 µM)
1 µL Template	0.2 µL Firepol taq polymerase
	1.5 µL Template
95°C for 3 minutes	
Followed by 29 cycles of:	95°C for 3 minutes
95°C for 40 seconds	Followed by 29 cycles of:
52°C for 40 seconds	95°C for 40 seconds
72°C for 2 minutes	52°C for 40 seconds
Then:	72°C for 2 minutes
72°C for 7 minutes	Then:
4°C hold	72°C for 7 minutes
	4°C hold
AcuPrime taq recipe	**Colony PCR**
20.3 µL H_2O	15.3 µL Sterile water
2.5 µL Buffer	2 µL NEB buffer
0.5 µL Forward primer (20 µM)	0.5 µL dNTPs
0.5 µL Reverse primer (20 µM)	0.5 µL M13 forward primer (20 µM)
0.2 µL AcuPrime taq polymerase	0.5 µL M13 reverse primer (20 µM)
1 µL Template	0.2 µL NEB taq polymerase
94°C for 3 minutes	95°C for 3 minutes
Followed by 29 cycles of:	Followed by 29 cycles of:
94°C for 30 seconds	95°C for 40 seconds
52°C for 30 seconds	52°C for 40 seconds
68°C for 1 minutes	72°C for 2 minutes
Then:	Then:
68°C for 10 minutes	72°C for 7 minutes
4°C hold	4°C hold

Sequencing was later performed according to a more efficient protocol modified by Fronny Plume for use in the LoST Lab (*Table 4*). Reactions were prepared in a 96 well plate. After the sequencing reaction, samples were prepared for capillary separation using a salt solution. A master mix of 1.95 ml 95% ethanol, 37.5 µL sterile water, 37.5 µL 0.25M EDTA and 150 µL 3M NaOAc was prepared for 48 sequences. 30 µL master mix was added to each reaction, and samples were mixed

then incubated at room temperature for ten minutes. Samples were then centrifuged at 3000rpm for 20 minutes at room temperature. Supernatant was discarded by inverting samples on a paper towel. 60 µL 70% ethanol was added to each sample, which were then centrifuged at 3000rpm for 20 minutes at room temperature. Supernatant was discarded by inverting samples on a paper towel. The plates were then centrifuged upside down on a paper towel at 750rpm for two minutes. 30 µL of 0.1mM EDTA was added to each sample. The plate was vortexed at moderate speed for 5 minutes. The plate was centrifuged at 700 rpm for one minute then submitted for capillary separation and analysis by the Massey Genome Service.

Table 4 Standard sequencing reaction protocols

LoSTLab Sanger Sequencing	Doyle Lab Sanger Sequencing
2.8 µL 5x ABI Sequencing Buffer	1-8 µL Cleaned PCR Template
0.8 µL ABI BigDye V3.1 Ready Reaction Mix	1 µL Primer (10 µM)
0.8 µL Primer (5 µM)	Sterile water, if necessary, up to 8.5 µL
3-10 µL Cleaned PCR Template	
Sterile water, if necessary, up to 16 µL	A master mix for 48 sequences:
	135 µL 5x ABI Sequencing Buffer
26 cycles of:	27 µL ABI BigDye V3.1 Ready Reaction Mix
96°C for 10 seconds	27 µL Sterile Water
50°C for 5 seconds	
60°C for 4 minutes	3.5 µL master mix added to each sample
4°C hold	
	96°C for 4 minutes
	Followed by 25 cycles of:
	96°C for 10 seconds
	57°C for 5 seconds
	60°C for 3 minutes
	10°C hold

While undertaking sequencing of our data set, the 1001 Genomes Project released several sets of 'resequenced genomes' from *A. thaliana* accessions (www.1001genomes.org). We acquired the 5' end of the sequence for (*AtMYB90*) *PAP2* in accession Sakata from this project, as we were not able to do so via Sanger sequencing methods. The partial sequence was acquired from: http://signal.salk.edu/atg1001/3.0/ge-browser.php. We did not analyse the resequence data

alongside our dataset of 48 accessions, as previous analyses have determined this inadvisable (Bloomer *et al.*, 2012) due to an apparent high error rate in the resequencing effort.

3.2.3 Gene Cloning. PCR reactions which resulted in multiple bands were cloned using Invitrogen PCR4-TOPO TA cloning protocol. Plates were prepared using autoclaved LB/agar with 50mg/L kanamycin. PCR templates were prepared by adding one unit of NEB taq polymerase and incubating at 72°C for 20 minutes to add adenine nucleobases to product ends. 1.33 µL PCR product was prepared for cloning with 0.33 µL Salt solution and 0.33 µL TOPO vector. The samples were mixed then left to incubate at room temperature for 30 minutes. 16 µL of TOP-10 *E.coli* cloning cells were added to each ligation and these were placed on ice for 30 minutes. The samples were then heated shocked at 45°C for exactly 45 seconds and placed immediately back on ice for five minutes. 80 µL SOC medium was added to each sample, which were then placed on a shaker at 37°C overnight. The samples were plated out at two different volumes: ten µL and 80 µL. The plates were then placed upside down at 37°C overnight.

Plates with successful colony growth had 9 colonies sampled by touching a sterile pipette tip to the colony and stirring in a prepared NEB PCR solution, using M13 forward and reverse primers (*Table 3*). Each of the samples was run on a 1% agarose gel, stained in ethidium bromide and visualised in gel dock under UV light. Samples that showed the expected band size after visualisation were sequenced using Sanger sequencing protocols and T3 forward primer and T7 reverse primers (*Table 4*).

The first intron of PAP2 has both a Poly-A sequence and Poly-AT sequence, rendering it virtually impossible to resolve the sequence between these regions using traditional sequencing methods. Instead, plasmid preparations from the positive colonies were used as sequencing template. Colonies were screened using NEB PCR and then visualised on a 1% agarose gel with ethidium bromide and UV light. The colonies showing PCR fragments of the expected size were cultured in 15ml falcon tubes in 3ml of Liquid Luria Broth (LB) media with 3µL of 50mg/ml kanamycin for 14 to 16 hours at 37°C. Spin Miniprep kit (50) (QIAGEN QIAprep®) was used to isolate plasmids for

sequencing. The cultures were pelleted in microcentrifuge tubes 1ml at a time (1ml pelleted, supernatant discarded; repeated until the whole culture was pelleted) at 8krpm for 3 minutes. The bacterial pellets were resuspended in 250 µL buffer P1. 250 µL buffer P2 was added and samples inverted 5 times, resulting in a clear solution. 350 µL buffer N3 was immediately added and samples inverted 5 times. Samples were centrifuged for 10 minutes at 13krpm. Supernatant was poured into spin columns and centrifuged for 45 seconds at 8krpm. Flow-through was discarded. Spin columns were washed with 500 µL buffer PB, centrifuged for 45 seconds at 8krpm. Flow-through was discarded. Spin columns were washed using 750 µL buffer PE, centrifuged for 45 seconds at 8krpm. Flow-through was discarded. Samples were centrifuged for a further 60 seconds at 8krpm. Spin columns were transferred to new microcentrifuge tubes. DNA was eluted by adding 50 µL buffer EB to the spin column, stood for 1 minute and then centrifuged for 1 minute at 8krpm. DNA samples were checked with M13 primers to ensure the desired insert had been cloned. Samples were quantified using NanoDrop ND-1000. 500-600ng plasmid DNA was used for each sequencing reaction. Standard Sanger sequencing was used (*Table 4*).

3.2.4 Molecular data analysis. Contiguous sequences for each gene were first assembled for individual accessions using MUSCLE in GENEIOUS v5.3.4 (Drummond *et al*., 2010) using the sequence from Col-0 as a template, acquired from TAIR (Lamesch *et al*., 2010). Using the start/stop boundaries from Columbia sequence, alignments of *A. thaliana* accessions were generated for each PAP gene and WER separately. Due to PCR and sequencing issues, genomic sequences could not be obtained for all accessions of each gene. We could only obtain coding region sequences for Ms-0, N7, Sav-0 and Yo-0 of PAP2 and Sakata of *AtMYB114* (*PAP4*). We could not sequence a complete sequence of the coding region of PAP2 from the accession Sakata; the first 646 bp of sequence (exon 1 and intron 1) were produced by the Weigel laboratory at the Max Planck Institute for Developmental Biology. To assess nucleotide diversity of our genomic alignments, a gene-wide measure of π was calculated as well as a sliding window analysis of π using a window of 40 bp and a step size of 3 bp in DNASP

v5.10 (Librado & Rozas, 2009). We analysed the four PAP genes, two predominant haplogroups of *AtMYB75* (*PAP1*), and *WER* separately.

Intron/exon boundaries from Columbia cDNA sequence were used to produce inferred cDNA alignments for each of the *PAP* and *WER* genes. In our analysis of *PAP4*, we found accession N6 had a deletion which compromised the splice site between exon 1 and intron 1. To infer likely splice sites and resulting peptides, we used GENSCAN (Burge & Karlin, 1997) using the *Arabidopsis* model and the default suboptimal exon cut-off of 1.00. The predicted peptide was used in subsequent analyses. Also in *PAP4,* we observed a single nucleotide polymorphism (SNP) in 10 accessions that results in an early stop codon; this SNP was described in PAP4 from Col-0 previously (Gonzalez *et al.*, 2008). Beyond the early stop codon in these 10 accessions, there are no other SNPs relative to the most common (presumably functional) *PAP4* haplotype. As such, we treated the stop codon of the other 38 accessions in our *PAP4* dataset as the end of genomic and cDNA sequences in the 10 accessions with the novel stop codon for the purposes of phylogenetic analysis, *de novo* motif identification and sequence comparisons between the *PAP* genes. Molecular analysis of *PAP4* using DNASP v5.10 (Librado & Rozas, 2009) treated the early stop codon as a rare allele.

Haplotype files for each cDNA alignment were generated using DNASP v5.10 (Librado & Rozas, 2009). The files were generated with gaps considered and exported to NETWORK v4.6 (Fluxus Engineering) to create haplotype networks using median joining (Bandelt *et al.*, 1999). Further to this, departure from a neutral pattern of evolution was tested by comparing alleles of the two haplogroups of *PAP1*, as well as between the four *PAP* genes and *WER*. To this end, a sliding window analysis of *Ka/Ks* ratios was performed across pair-wise cDNA alignments of the four *PAP* genes, with a window of 40bp and a step size of 3bp using DNASP v5.10 (Librado & Rozas, 2009). Where *Ks* values of 0 occurred, due to very low polymorphism in regions of the genes, a ratio could not be calculated. To overcome this, sliding window values for *Ka* only were used and the ratio was calculated using a gene-wide value for *Ks* as in Shiu *et al.* (2004).

The four *PAP* loci are clearly the result of duplication events. In an attempt to resolve the duplication history at these genes, a Bayesian Analysis was performed using MRBAYES v3.2 (Ronquist & Huelsenbeck, 2003). *AtMYB82* was used as an outgroup, as it is a member of the *A. thaliana* R2R3 MYB gene family (Stracke *et al*., 2001), and therefore related to the PAP gene family as a whole, but unrelated to any of the PAPs sufficient to potentially allow resolution of the PAP gene family phylogeny. We used the default evolutionary model (GTR substitution model with gamma-distributed rate variation across sites and a proportion of invariable sites). Diagnostics were calculated every 1000 generations, the sample frequency was calculated every 100 generations and the print frequency was every 1000 generations; the analysis was halted once the standard deviation of split frequencies fell below 0.01, as recommended by the program manual. Tree files were visualised and rooted using FIGTREE v1.4 (tree.bio.ed.ac.uk/software/figtree). Linkage disequilibrium analyses were performed by concatenating the four genes for each accession in the order they occur on the chromosome (*PAP1:PAP3:PAP4:PAP2*) in GENEIOUS; our dataset allowed 45 complete concatenated sequences. The concatenated sequences were then aligned using the "muscle" option and the resulting FASTA data file was then opened in TASSEL v4.0 (Bradbury *et al*., 2007), the dataset was filtered to only consider variable sites and a full matrix linkage disequilibrium analysis was performed. MEME-CHIP (Machanick & Bailey, 2011) was used in an attempt to identify novel motifs useful for identification of genes of the R2R3-MYB gene family and to further elucidate the evolutionary history of closely related genes with in this family. We also used the protein structure prediction software of GENIOUS to suggest likely protein stages resulting from mutational changes to gene sequences. Using inferred cDNA sequences of the PAP genes, we used the NCBI BLASTN SUITE (http://blast.ncbi.nlm.nih.gov, Zhang *et al.*, 2000) to identify orthologs of the PAPs in the related species *Brassica rapa* and *Arabidopsis lyrata*.

3.3 Results

3.3.1 Nucleotide diversity and patterns of polymorphism in genomic alignments of the PAP and WER loci

3.3.1.1 PAP1. Gene-wide, the nucleotide diversity (π) for PAP1 from the 48 accessions surveyed analysed was 0.00963. A sliding window analysis of π identified a single region of low variation over the first exon and the beginning of the first intron (*Figure 3*). However, beyond this, 15 bp into the first intron, variation increases suddenly and fluctuates for the remainder of the gene. The only area of the gene with comparable levels of low variation was located at the 3' end of the gene; no variation whatsoever was observed in the final 70 bases of the alignment. Towards the centre of the gene we did observe a region of comparatively low variation corresponding to the end of the second exon and the start of the second intron. Still, the variation observed here was still higher than that observed at the beginning and end of the *PAP1* genomic alignment.

A. thaliana apparently maintains two alleles of *PAP1* based on 35 high frequency single nucleotide polymorphisms (SNPs), seven in the coding regions, 28 in the introns. Of the seven coding region SNPs, six are nonsynonymous. As most of the nucleotide diversity we observed was high frequency and either present or absent for a haplogroup, we analysed π for the two allelic groups separately:

3.3.1.1.1 PAP1 haplogroup A. As expected, the gene-wide nucleotide diversity of the most common haplogroup of *PAP1* (P1A), present in 39 accessions in our dataset, was much lower than that of the complete *PAP1* genomic alignment, at 0.00152. A sliding window analysis of π identified a peak of diversity towards the start of intron 1, evidently the result of a number of indels in this area; this peak was previously observed in the complete *PAP1* genomic alignment (*Figure 4*). π is markedly reduced for the remainder of the alignment in this haplogroup, supporting the notion that the majority of nucleotide diversity in the *PAP1* alignment is the result of two diverged haplogroups.

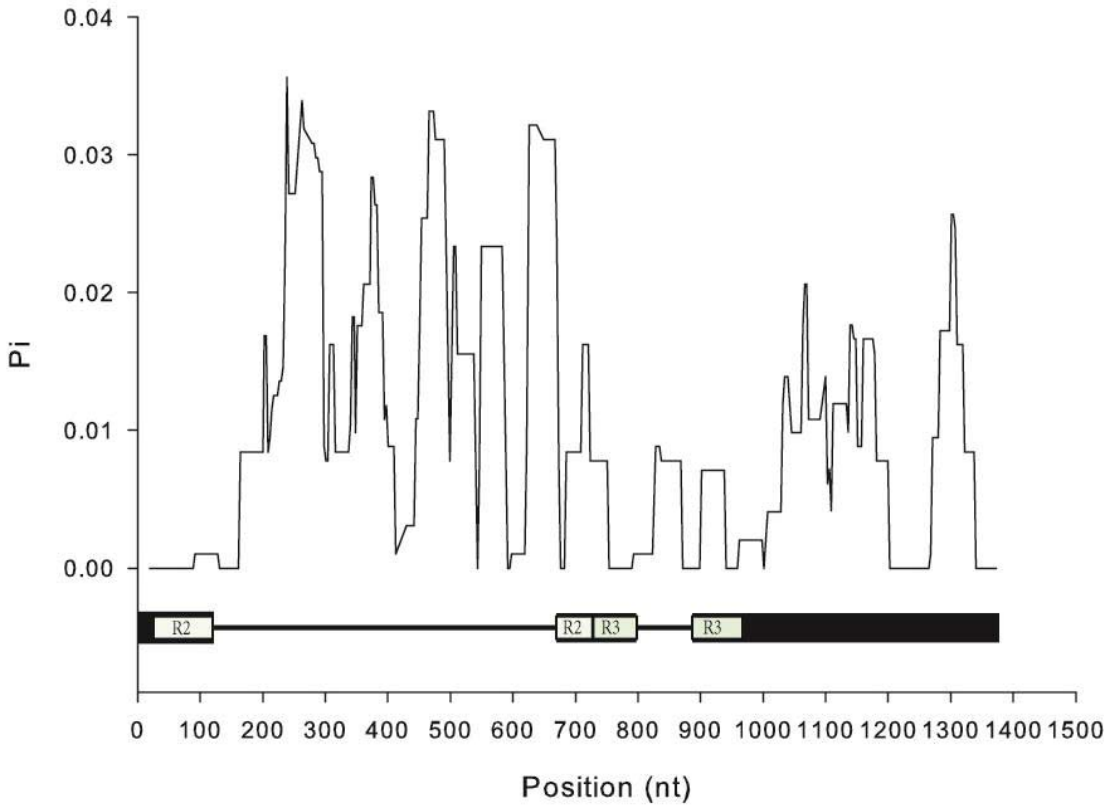


Figure 3 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP1* genomic sequences from 48 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP1* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

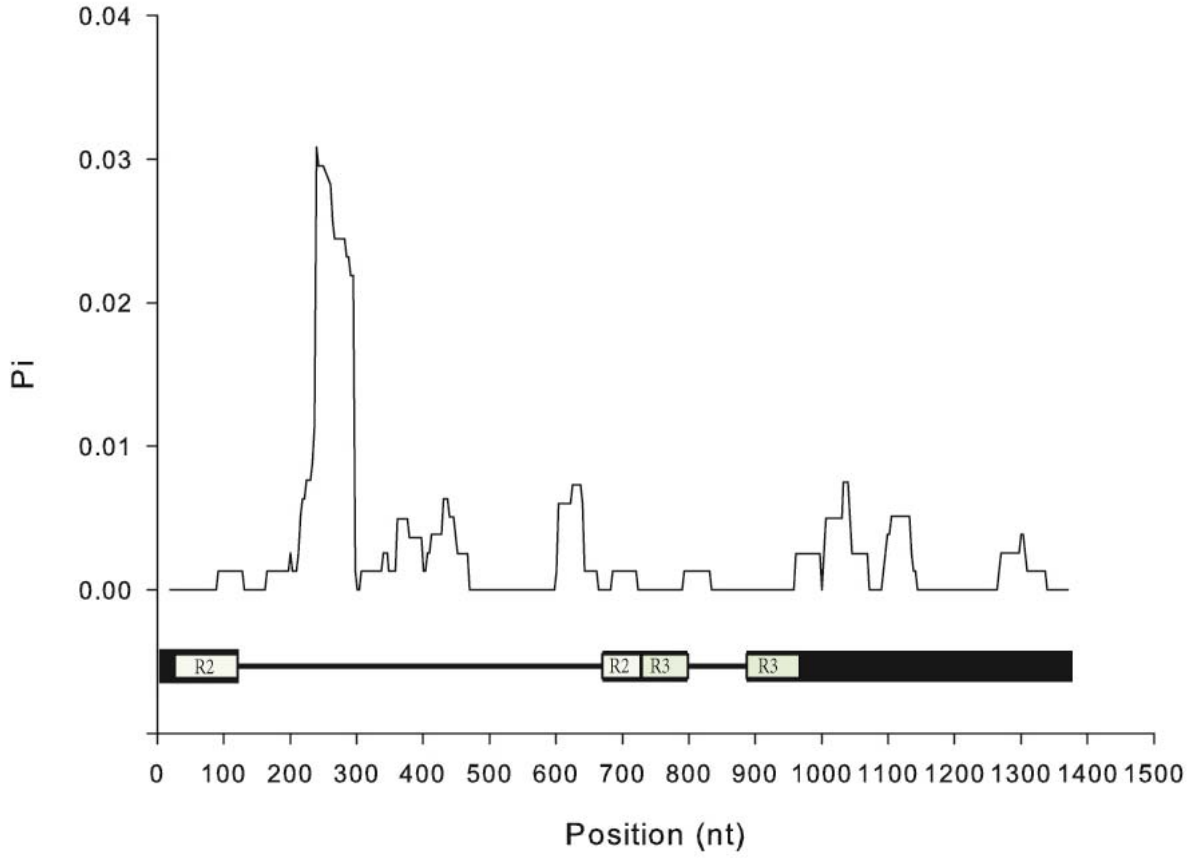


Figure 4 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP1* genomic sequences comprising the P1A haplogroup from 39 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP1* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

3.3.1.1.2 PAP1 haplogroup B. The less frequently sampled *PAP1* haplogroup (P1B), found in nine accessions, has a lower level of gene-wide π than that of P1A, at 0.00133. Using a sliding window analysis of nucleotide diversity, the alignment revealed five peaks of nucleotide diversity (*Figure 5*). The first two peaks were located in second half of intron 1 corresponding to nucleotide replacements and deletions of sequence; we saw no nucleotide diversity upstream of this, in exon 1 or the first half of intron 1. Exon 2 and intron 2 also revealed no nucleotide diversity, with the last three sites located in exon 3: S179N, I215M and A220T (Here and herein, the amino acid to the left of the number indicates the most common amino acid in the alignment; the amino acid to the right indicates the amino acid found in the subject referred to).

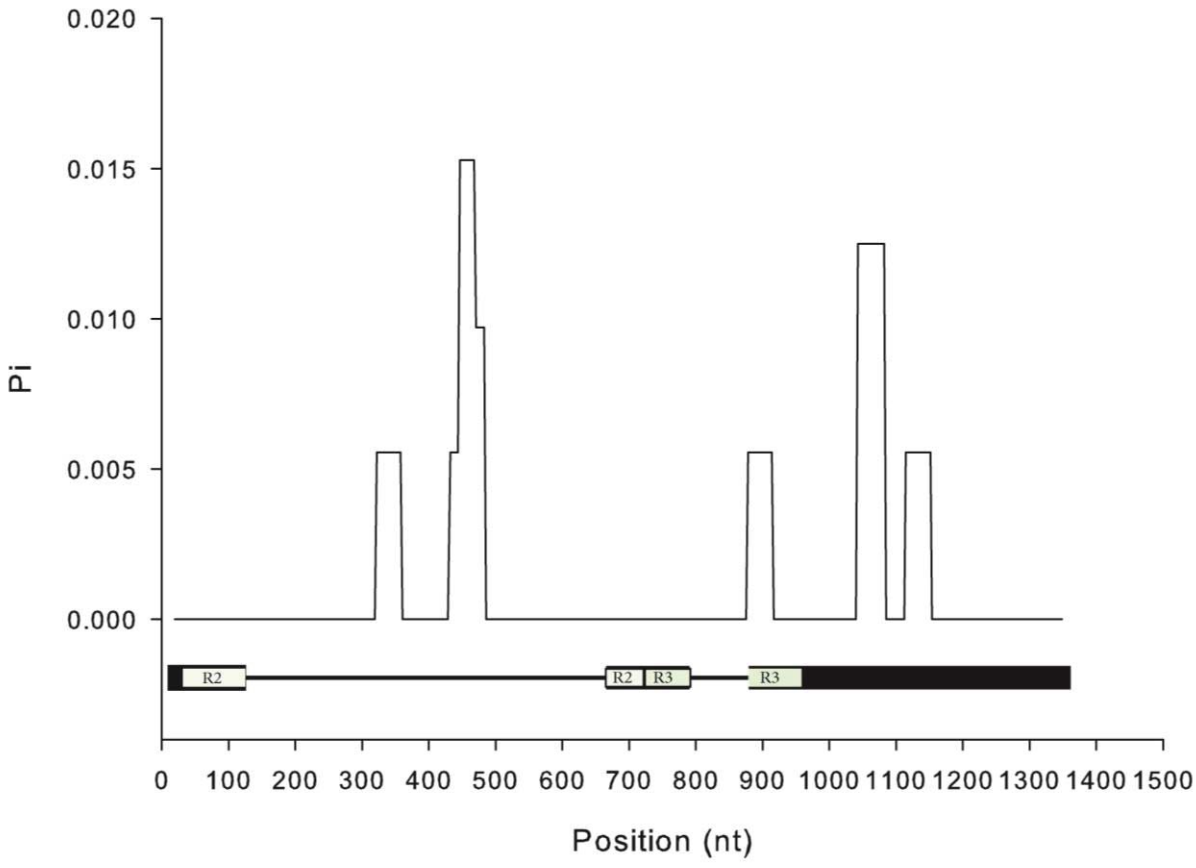


Figure 5 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP1* genomic sequences comprising the P1B haplogroup from nine accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP1* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

3.3.1.2 PAP2. The gene-wide nucleotide diversity of the 43 accessions of PAP2 surveyed was 0.00277. A sliding window analysis of π across the alignment revealed no variation through exon 1, though nucleotide diversity of PAP2 peaks almost immediately into intron 1, the result of several high frequency nucleotide replacements as well as a high frequency poly-AT insertion (*Figure 6*). However, these sequence variations are confined to the first 200 bp of intron 1, less than half the total length of the intron. The remainder of the alignment reveals little variation; two large peaks of variation towards the end of exon 3 correspond to high frequency nucleotide substitutions, while sites of moderate variation can be attributed to discrete nucleotide replacements in single accessions. As was seen previously in *PAP1*, we observed no nucleotide variation towards the 3' end of the *PAP2* genomic alignment.

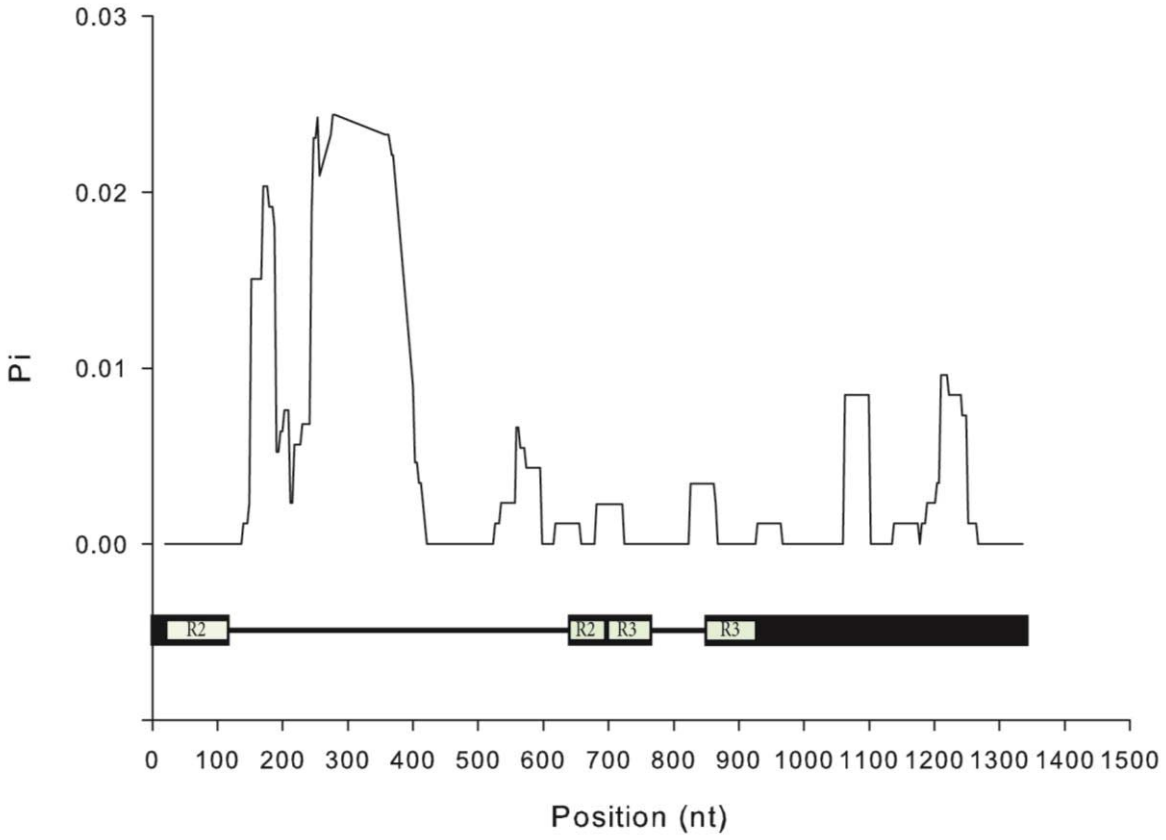


Figure 6 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP2* genomic sequences from 38 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP2* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

3.3.1.3 PAP3. Gene-wide, the measure of π for the 47 accessions of *PAP3* surveyed was 0.00252. However, unlike *PAP1* and *PAP2*, a sliding window analysis of π revealed nucleotide diversity to be highest in exon 1, the result of several discrete mutations and a single high frequency polymorphism (*Figure 7*). Conversely, the first half of intron 1 is the region with the lowest variance of π of the whole PAP3 alignment; only moderate nucleotide variation (π < 0.01) was observed in the last 40 bp of intron 1. Interestingly, the length of intron 1 in *PAP3* is by far the shortest of the PAP genes, only 145 bp long while intron 1 of *PAP1, PAP2* and *PAP4* exceeds 500 bp in length. A second peak of π was observed over the majority of intron 2, the result of a high frequency polymorphism and a number of discrete SNPs.

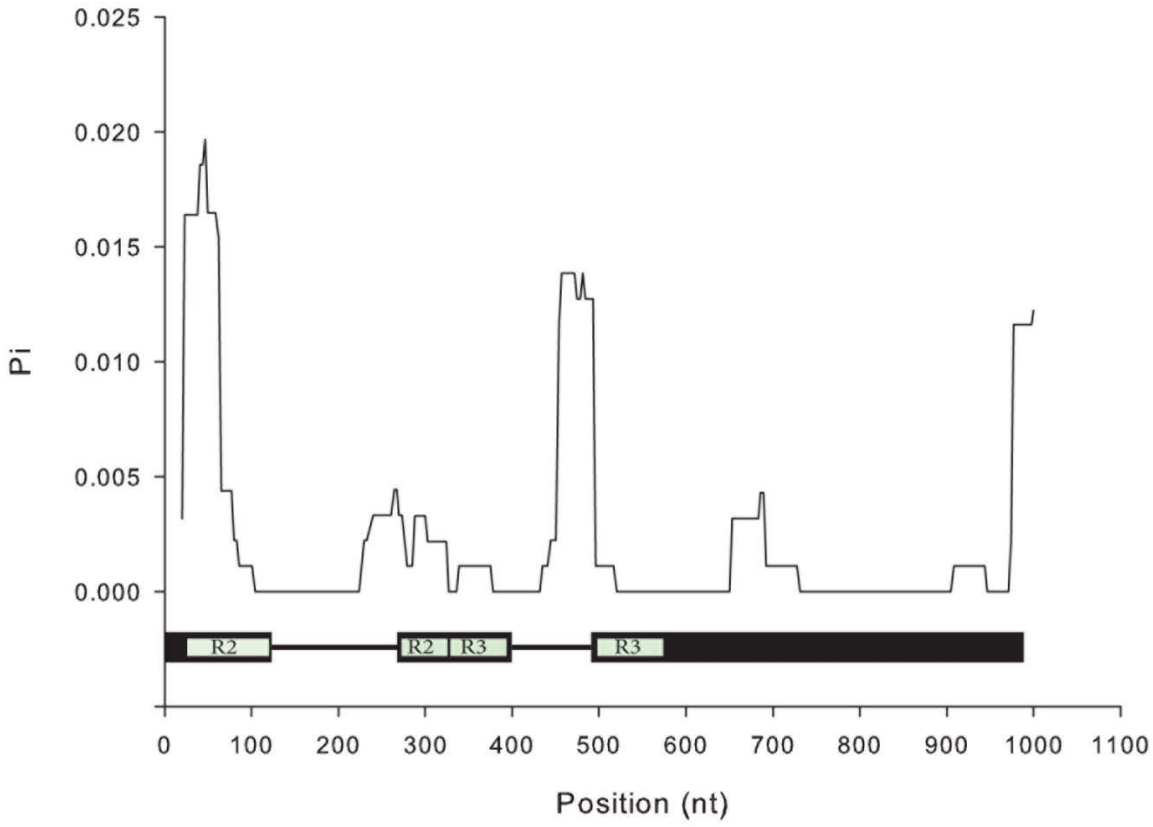


Figure 7 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP3* genomic sequences from 37 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP3* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

Unlike *PAP1* and *PAP2*, we observed nucleotide variation at the 3' end of the alignment due to a high frequency nucleotide replacement resulting in amino acid change D239E.

3.3.1.4 PAP4. The gene-wide measure of π of the 47 accessions of *PAP4* surveyed was 0.00686. As was observed in *PAP3*, some nucleotide variation was revealed in exon 1 by a sliding window analysis of π, though the level of variation was less than that observed in exon 1 of *PAP3* (*Figure 8*). Relative to the rest of the *PAP4* alignment, the region of the gene including exon 1 and first half of intron 2 revealed the least nucleotide diversity, punctuated by a single peak of variation corresponding to a high frequency SNP in close proximity to a high frequency 21 bp indel mutation. The largest peak of π covered the region of the gene from the centre of intron 1 to the end of the intron. Underlying this peak were several high frequency and discrete SNPs and indels. No nucleotide variation was observed in the regions corresponding to exon 2 or the first half of exon 3. Given its short length, intron 2 exhibited a surprisingly high level of nucleotide diversity, with all five variable sites found being of high frequency and found in the same 10 accessions; levels of variation in intron 2 of *PAP4* exceeded peaks of nucleotide variation at any location throughout the other three *PAP* genes. We observed a peak of variation towards the start of exon 3; two high frequency SNPs found in two discrete groups of accessions underlie this peak. A second, smaller peak, towards the middle of exon 2, is the result of a 12 bp insertion in three accessions. As in *PAP3,* we observed nucleotide variation at the 3' end of the alignment in *PAP4*, though the level of π was moderately less than that of *PAP3* (π (*PAP3*) ≈ 0.015; π (*PAP4*) ≈ 0.01).

The genomic alignment of *PAP4* reveals 12 high frequency mutations that were maintained exclusively by 10 accessions, with another two high frequency mutations maintained by, but not exclusive to, this group (Bl-1, Bur-0, Mh-1, Ms-0, Nok-1, Pyl-1, Rld-2, Rubezhnoe-1, Sah-0, Sap-0). Interestingly, the majority of the polymorphisms maintained by this group of accessions are located in the introns (10 of 14); the five high frequency polymorphisms found in intron 2 were exclusive to this group and contribute to this division. Half of these unique polymorphisms are located in the 71

bp long intron 2 (84 bp with insertion included), including a 13 bp insertion. Another high frequency insertion of 21 bp was found towards the beginning of intron 1, though this mutation is also maintained by three other accessions as well.

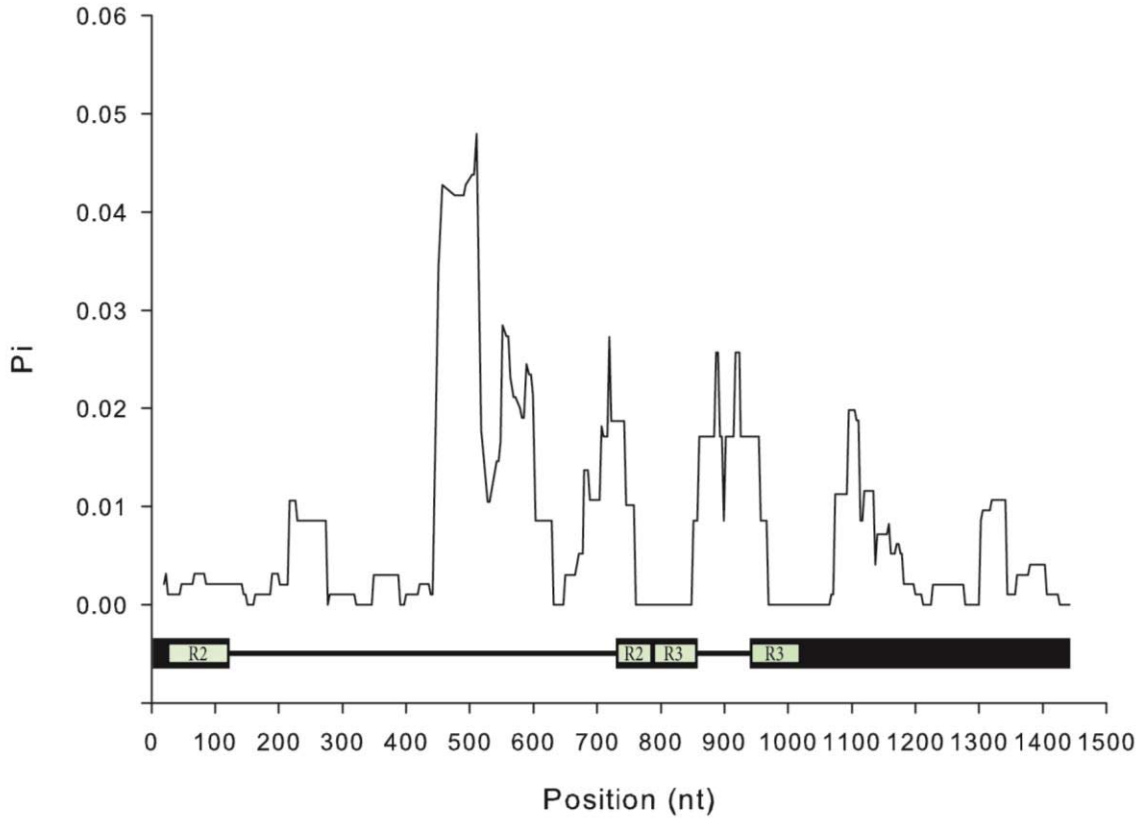


Figure 8 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *PAP4* genomic sequences from 47 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *PAP4* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

3.3.1.5 WER. Gene-wide, variance of π for the 48 accessions of *WER* surveyed was by far the lowest at 0.00092. A sliding window analysis of π revealed low nucleotide diversity across Exon 1, whereas intron 1 has a peak of high nucleotide diversity compared to the rest of the alignment, the result of a high frequency insertion mutation and a SNP towards the 5' end of intron 1 (*Figure 9*). The rest of the alignment reveals very little variation, except for a peak (π = 0.012) straddling the border of exon 2 and intron 2, the result of several low frequency SNPs in close proximity.

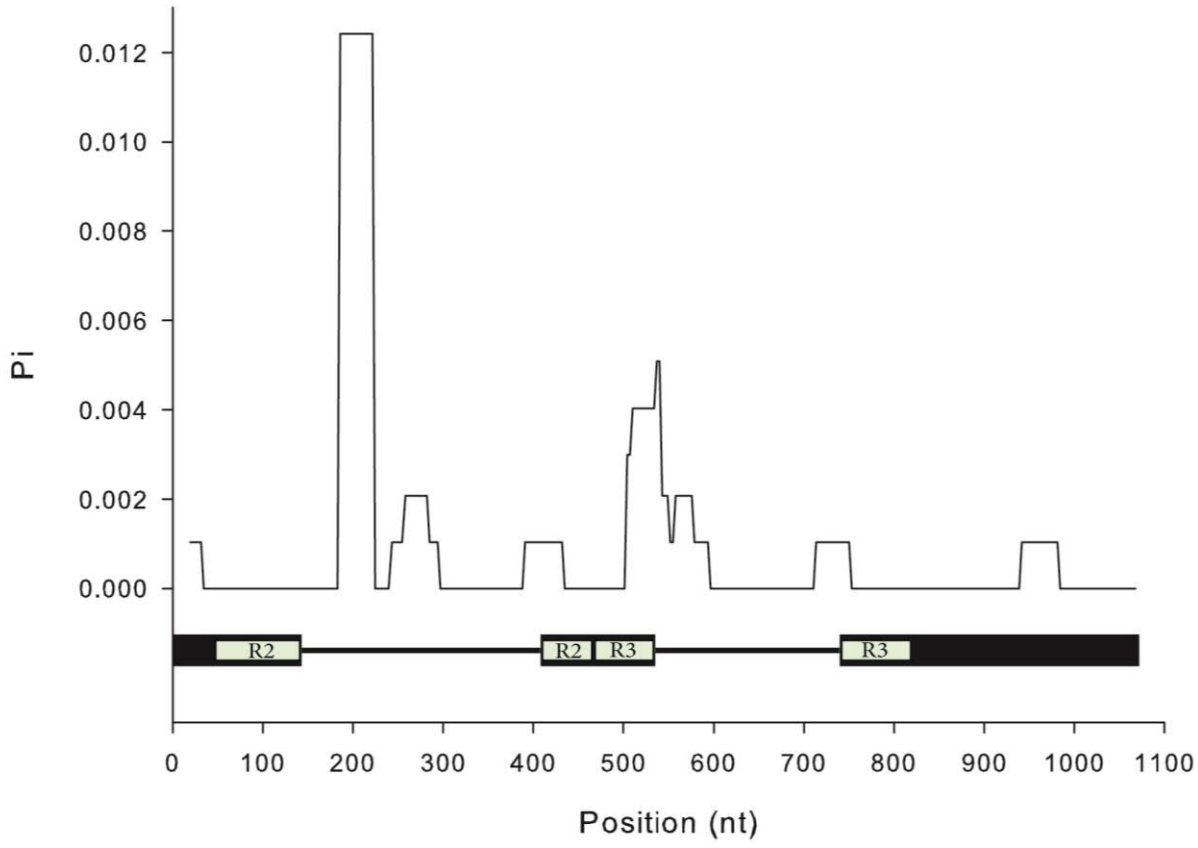


Figure 9 Sliding window analysis of nucleotide diversity (Pi) for an alignment of *WER* genomic sequences from 48 accessions of *Arabidopsis thaliana* with Pi plotted against window midpoint. The underlying schematic indicates positions of the three *WER* exons (black boxes), introns (black lines) with positions of the R2R3 MYB domains indicated by underlying white boxes.

3.3.2 Intragenic variation of the coding regions

3.3.2.1 PAP1. Based haplotype network data, we identified eight haplotypes for PAP1 (*Figure 10*). The haplotypes are split into two haplogroups based on seven high frequency single nucleotide

polymorphisms (SNPs), six of which are nonsynonymous (*Table 5*). P1A includes 39 of the 48 analysed accessions and five haplotypes, with the most frequently sampled haplotype for *PAP1*, represented by 33 accessions. P1B includes three haplotypes sampled from nine accessions in our dataset, differentiated by three low frequency SNPs, two of which are nonsynonymous. Initially, the accessions Sakata seemed to give conflicting results during analysis, as it shared some of the mutations separating P1A and P1B, but not all. Assuming this to be editing or sequencing error, repeat sequencing was undertaken as well as cross-checking our sequence with that produced by the 1001 genome project (www.1001genomes.org), though all efforts still suggested that the accession Sakata does only share three polymorphisms with other accessions belonging to P1B, as well maintaining four unique SNPs, two of which are non-synonymous. In this case, it appears the *PAP1* gene has undergone recombination in the past, and for this we excluded Sakata from either predominant haplogroup. We observed a concentration of polymorphisms immediately outside of the conserved R2R3 MYB region (11 of a total 13 non-synonymous SNPs); 69% of the nonsynonymous mutations are found in an area of the gene accounting for only 17% of the coding region, between the MYB region and the final 70 amino acids of the coding sequence (*Figure 11*).

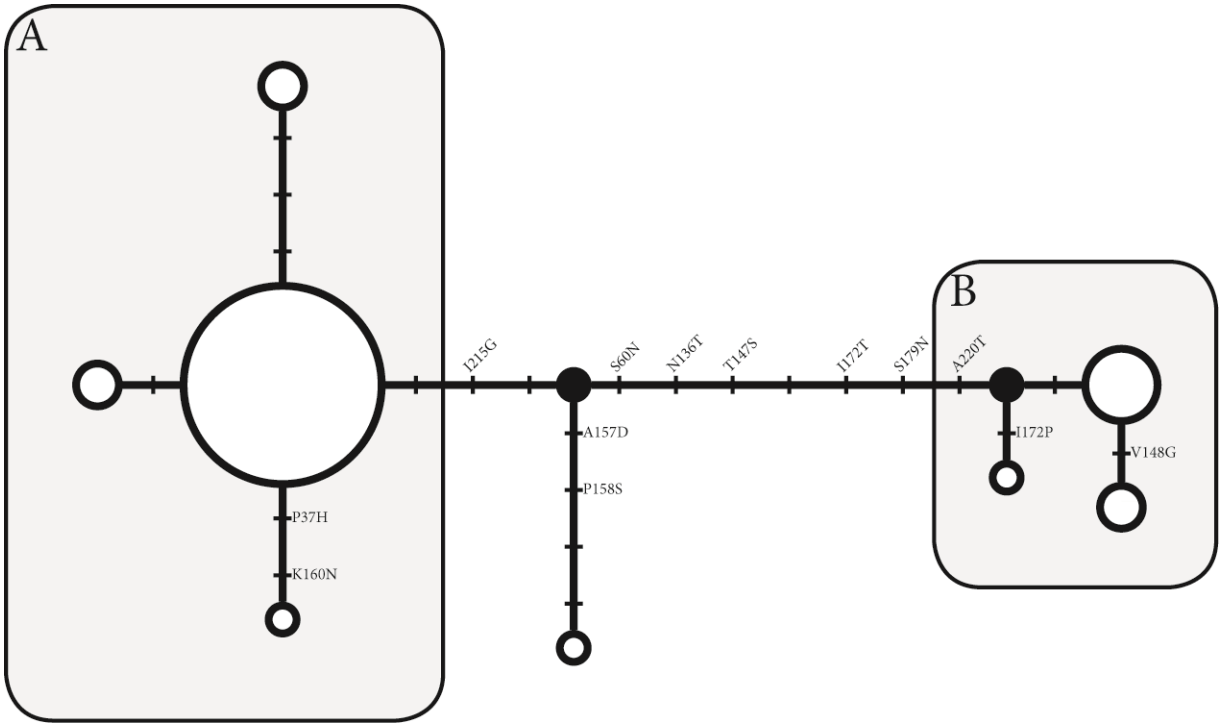


Figure 10 Median-joining haplotype network of *PAP1* coding region alleles. Eight haplotypes were identified based on inferred cDNA nucleotide sequence from 48 *Arabidopsis thaliana* natural accessions; haplotypes are represented by open circles, and their frequencies indicated by relative circle size. The black-filled circles represent hypothetical, unsampled haplotypes required to complete the network. Branch lengths and tick marks reflect the number of nucleotide changes separating haplotypes. Alleles belonging to haplotypes A and B are circumscribed by shaded boxes labelled A and B.

Table 5 *PAP* amino acid replacements and indels identified from 48 *Arabidopsis thaliana* accessions

Protein region	*PAP1**	*PAP2**	*PAP3**	*PAP4**	Accessions carrying minority allele
'Undefined' domain				E2V	Ita-0
			G7W		Sah-0
R2 MYB domain			K10R		Bl-1, Ms-0, Rld-2, Sap-0
				T14S	Ge-0
			T15A		Akita, Alc-0, Bl-1, Edi-0, Ge-0, Jea, Ms-0, N6, Oy-0, Pi-0, Ri-0, Rld-2, Rubezhnoe-1, Sah-0, Sap-0, Te-0, Tsu-0
			T15S		Can-0
			R22S		Akita
				R22K	Cvi-0
		L23Q			N7, Sav-0
		D26G			N7, Sav-0
				K27N	N14
				KWHQVPLRA32-41R	N6
				Q35E	Alc-0, Kondara, N13, N14, Shahdara, Ta-0
	P37H				Cvi-0
	S60N[Ŧ]				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0
R3 MYB domain			81PolyA		Sah-0
			G96R		Sah-0
			N116K (FS)		Rld-2
'Undefined' domain (*cont.*)			R129L		Sah-0
				N130K	Sah-0
				P132L	Bl-1, Bur-0, Cvi-0, Ita-0, Mh-1, Ms-0, Nok-1, Pyl-1, Rld-2, Rubezhnoe-1, Sah-0, Sap-0, Te-0
	N136T[Ŧ]				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0
'Subgroup 6' motif				K140STOP	Enkheim-T, Gre-0, Jm-0, Mt-0, N7, Oy-0, Pa-1, Sav-0, Sp-0, Yo-0
'Undefined' domain (*cont.*)	T147S[Ŧ]				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0

Protein Region (*cont.)*	PAP1* (*cont.*)	PAP2* (*cont.*)	PAP3* (*cont.*)	PAP4* (*cont.*)	Accessions carrying minority allele (*cont.*)
'Undefined' domain (*cont.*)	V148G				Bur-0, N7, N13
				I148L	Can-0, Edi-0, Pi-0
			V153I		Sah-0
				N153K	N14
				L155F	Bur-0, Nok-1, Pyl-1
			PR156-157#		Alc-0, Bl-1, Ms-0, Rld-2, Sap-0
			R156-157K#		Bur-0, Nok-1
	A157D				Sakata
	P158S				Sakata
			R159G#		Alc-0
	K160N				Cvi-0
				V161L	Cvi-0
				P165S	Sah-0
	I172P				Ishikawa
	I172T‡				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0
			I172V		Sah-0
	S179N‡				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0
			H182R		Sah-0
				N183S	Ct-1, Stw-0
				184-187KDDE	Can-0, Edi-0, Pi-0
		N184K			Kondara
			Y189N		Sah-0
		G199E			Ita-0
		N201K			Rld-2
		E209G			Enkheim-T, Gre-0, Jm-0, Mt-0, N7, Oy-0, Pa-1, Sav-0, Yo-0
				R205G	Bl-1, Bur-0, Mh-1, Ms-1, Nok-1, Pyl-1, Rld-2, Rubezhnoe-1, Sah-0, Sap-0
				R205P	Jea
				G210D	Cvi-0
	I215M‡				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0, Sakata

Protein region (*cont.*)	*PAP1** (cont.)	*PAP2** (*cont.*)	*PAP3** (*cont.*)	*PAP4** (*cont.*)	Accessions carrying minority allele (*cont.*)
'Undefined' domain (*cont.*)	A220T[Ŧ]				Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0
				E224K	Can-0, Edi-0, Pi-0
				228L (FS)	Ita-0
			W236C		Sah-0
			F239G		Kondara, Shahdara
			D240E		Bla-1, Blh-1, Cvi-0, Ishikawa, Kondara, Lip-0, Mh-1, Ran, Sakata, Shahdara, Ta-0

*The minority allele is shown to the right of the amino acid position

[Ŧ]Polymorphisms which distinguish haplogroups A and B in *AtMYB75*

[#]These polymorphisms occur in the 'subgroup 6' motif in *AtMYB113,* though the position is further downstream than the others due to the polyA insertion upstream

Haplogroups of *AtMYB75*

Haplogroup	Accessions
A	Akita, Alc-0, Bla-1, Blh-1, Can-0, Ct-1, Cvi-0, Edi-0, Enkheim-T, Ge-0, Gre-0, Jea, Jm-0, Kondara, Kn-0, Lip-0, Mh-1, Mt-0, Ms-0, N6, Nok-1, Oy-0, Pa-1, Pi-0, Pyl-1, Ran, Ri-0, Rld-2, Rubezhnoe-1, Sakata, Sap-0, Sav-0, Shahdara, Sp-0, St-0, Stw-0, Ta-0, Tsu-0, Yo-0
B	Bl-1, Bur-0, Ishikawa, Ita-0, N7, N13, N14, Sah-0, Te-0

Gene-wide, nucleotide diversity of the inferred cDNA alignment of 48 accessions of PAP1 surveyed was 0.00569. The level of nucleotide diversity of synonymous sites is more than twice that of non-synonymous sites, at 0.01051 and 0.00439, respectively, indicating purifying selection may be acting on the *PAP1* locus. A summary of all nucleotide diversity values can be found in Table 6. Nucleotide diversity was also calculated across the whole alignment separately for each of the two haplogroups and was found to be somewhat lower in P1A (π = 0.00116) than in P1B (π = 0.00127). The difference in π between synonymous and non-synonymous sites for P1A ($\pi(s)$ = 0.00379; $\pi(a)$ = 0.00044) suggests negative purifying selection is acting on the haplogroup; on the other hand, the difference between $\pi(s)$ and $\pi(a)$ of P1B is much narrower (0.00137 and 0.00124, respectively). To

evaluate potential departure from neutral expectation for PAP1, Tajima's *D* and Fu and Li's *D** and *F** were calculated but were all found not to be significant.

Figure 11 Schematic representation of the PAP1 protein showing positions of amino acid replacements in 48 accessions. The horizontal black bar represents the full length protein, with the positions of the MYB domains (R2 and R3) shown as blue boxes. Each small vertical bar above the schematic indicates the position of an amino acid replacement table (Table 5). The seven linked replacements that define haplogroups A and B are shown with open squares at the top. A single small vertical bar sits below the full length protein schematic, as this occurs at the same site as a replacement associated with haplogroup definition in other accessions.

3.3.2.2 PAP2. The haplotype network generated for *PAP2* revealed eleven haplotypes, but no convincing evidence of haplogroup divisions as with *PAP1* (*Figure 12)*. The most common haplotype, from 24 accessions, is closely related to two less common haplotypes differentiated from the most common haplotype by a single SNP each; one from nine accessions by a synonymous SNP, the other from seven accessions by the nonsynonymous replacement E209G (*Table 5*). Gene-wide, nucleotide diversity of the inferred cDNA alignment of 48 accessions of *PAP2* surveyed was 0.00171, notably lower in comparison to *PAP1*. The measure of π(s) was much greater than that of π(a), at 0.00417 and 0.00103, respectively. Again, Tajima's *D* and Fu and Li's *D** and *F** were insignificant for *PAP2*, suggesting selective neutrality.

Table 6 A summary of measures of nucleotide diversity across the genomic and inferred coding sequences of the *PAP* and *WER* genes

Gene	π		cDNA		Tajima's *D*	Fu and Li's	
	gene-wide	cDNA	π(s)	π(a)		D*	F*
PAP1	0.00963	0.00569	0.01051	0.00439	NS	NS	NS
·P1A	0.00152	0.00116	0.00379	0.00044	NS	NS	NS
·P1B	0.00133	0.00127	0.00137	0.00124	NS	NS	NS
PAP2	0.00277	0.00171	0.00417	0.00103	NS	NS	NS
PAP3	0.00252	0.0023	0.00224	0.00235	NS	NS	NS
PAP4	0.00312	0.0039	0.00559	0.0035	NS	NS	NS
WER	0.00092	0.0004	0.00136	0.00017	NS	NS	NS

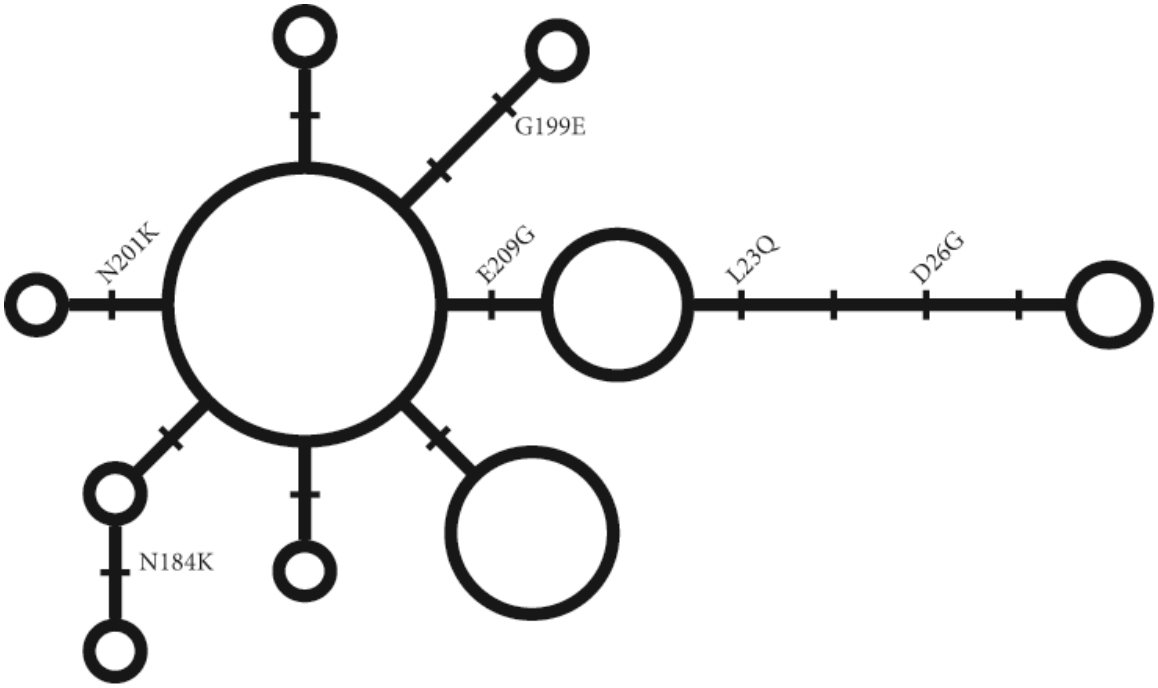


Figure 12 Median-joining haplotype network of *PAP2* coding region alleles. Ten haplotypes were identified based on inferred cDNA nucleotide sequence from 48 *Arabidopsis thaliana* natural accessions; haplotypes are represented by open circles, and their frequencies indicated by relative circle size. Branch lengths and tick marks reflect the number of nucleotide changes separating haplotypes.

The most genetically distant *PAP2* haplotype, represented by two accessions, is differentiated by five polymorphisms from most frequently sampled haplotype. The remaining six accessions, represented by one accession each, are differentiated by seven polymorphisms in total, three of which are nonsynonymous. Of the four PAP genes analysed, PAP2 has the fewest number of nonsynonymous mutations amongst the 48 accessions, with only six nonsynonymous SNPs in total. Two are found in the R2 repeat region, L23Q and D26G, shared by two accessions of a single haplotype (*Figure 13*). The other four are located in the undefined region downstream of the MYB regions. One of these four SNPs, E209G, is shared by nine accessions and two haplotypes. Two of these accessions, N7 and Sav-0, have E209G as well as the aforementioned SNPs L23Q and D26G, assigning these two to a distinct haplotype; the other seven accessions with E209G are a single haplotype. The other three SNPs are found in three separate accessions, N184K in Kondara, G199E in Ita-0 and N201K in Rld-2.

Figure 13 Schematic representation of the PAP2 protein showing positions of amino acid replacements in 48 accessions. The horizontal black bar represents the full length protein, with the positions of the MYB domains (R2 and R3) shown as blue boxes. Each small vertical bar above the schematic indicates the position of an amino acid replacement table (Table 5).

3.3.2.3 PAP3. Relative to *PAP1* and *PAP2*, the haplotype network of *PAP3* increases in complexity, revealing 14 haplotypes (*Figure 14*). Gene-wide, nucleotide diversity of the inferred cDNA alignment of 45 accessions of *PAP3* surveyed was 0.00230. The accessions Sah-0 and Rld-2 of PAP3 were excluded from analysis, as we propose their unique mutations to result in non-functional proteins and the mutations to artificially influence analysis (discussed later). Nucleotide diversity differences between π(s) and π(a) of *PAP3* was the inverse of that seen in *PAP1* and *PAP2*, at 0.00224 and 0.00235, respectively. Tajima's *D*, and Fu and Li's *D** and *F** were not found to be significant, again indicating neutral evolution.

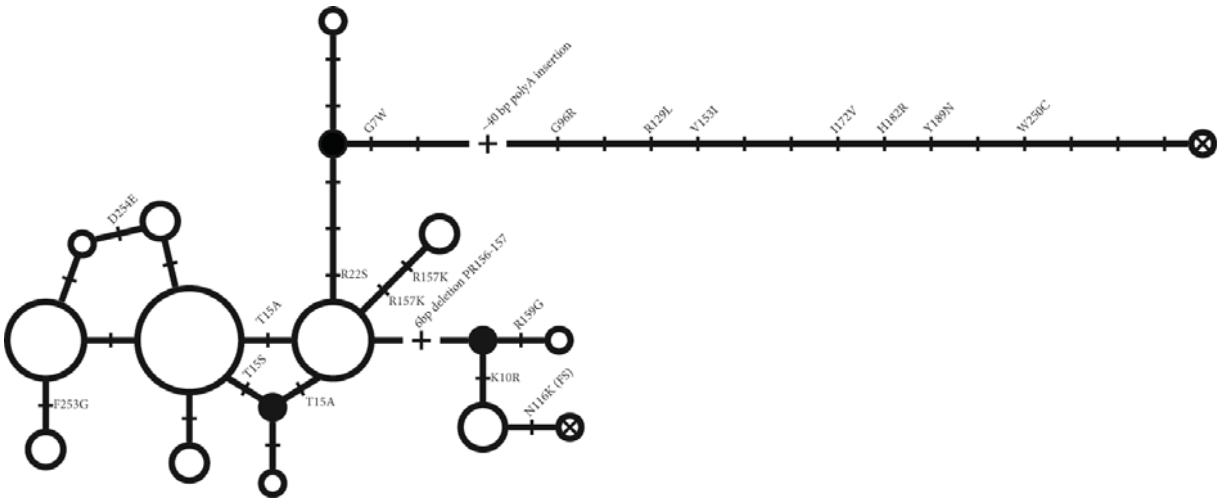


Figure 14 Median-joining haplotype network of *PAP3* coding region alleles. 14 haplotypes were identified based on inferred cDNA nucleotide sequence from 45 *Arabidopsis thaliana* natural accessions; haplotypes are represented by open circles, and their frequencies indicated by relative circle size. The black-filled circles represent hypothetical, unsampled haplotypes required to complete the network. The crossed circles represent putative dead alleles. Branch lengths and tick marks reflect the number of nucleotide changes separating haplotypes, with the exception of the dashed lines; these represent indels of varying lengths and are labelled accordingly.

The three most commonly sampled haplotypes are differentiated from each other by two mutations, one of which is the nonsynonymous replacement T15A (*Table 5*); in total, 30 accessions from our PAP3 dataset belong to these haplotypes. A subset of three haplotypes, to which five accessions from our dataset belong, is differentiated from the other haplotypes in the PAP3 network by an in-frame deletion of six base pairs (PR141-142--), as well as a hypothetical, unsampled haplotype required to complete the network. The deletion is located within the conserved 'subgroup 6' motif KPRPR[$^{S}/_{T}$]F proposed by Stracke *et al.* (2001). It is noteworthy that the two deleted residues, proline and leucine, are consecutively repeated in this region once. Other than this deletion, there are few mutations in the accessions which carry it: one haplotype, which only Alc-0 possesses from our dataset, has the mutation R145G downstream of the deletion; the other two haplotypes, of four accessions in total, share the nonsynonymous mutation K10R. Based on this, there is little evidence of mutational compensation in response to the deletion which these accessions share, though the expectation of such compensation is cursory given the deletion is in-frame and located outside of the highly conserved R2R3-MYB repeat regions.

We identified an allele defined by a frameshift mutation located in the R3 repeat region carried by a single accession, Rld-2, which also carries the in-frame deletion. The mutation, N102K, is resultant of an adenine nucleobase insertion. Given that this accession carries the aforementioned conserved deletion downstream of this frameshift mutation, it is likely that the former predates the latter. The deletion is found in the 'turn' region of the R3 'helix-turn-helix' secondary gene structure. The mutation results in eight out of frame residues, resulting in the third tryptophan residue in the R3 repeat region being replaced and an early stop codon one amino acid short of the end of the R3 region. Considering the importance of the regularly spaced tryptophan residues for the tertiary structure of the protein (Dubos *et al*., 2010), it is likely that the gene is rendered non-functional in this accession.

We identified a second frameshift mutation in PAP3, a poly-A insertion carried by accession Sah-0, at amino acid site 81, within the R3 repeat region. The insertion is roughly 40 base pairs long, though it is not possible to determine the precise number of base pairs with Sanger sequencing methods. Downstream of the mutation, 14 SNPs were found, seven of which are nonsynonymous. Upstream of the poly-A mutation, only two unique SNPs were found, one of which was nonsynonymous, resulting in the replacement of a glycine residue with a tryptophan residue at amino acid site seven. Only two mutations were identified upstream of the poly-A insertion, T15A, which is a high frequency mutation also found in *PAP3* of 16 other accessions in our dataset, and G7W, which is unique to the *PAP3* locus of Sah-0. We also noticed the distribution of SNPs in *PAP3* closely mirrored that observed in PAP1. 38% of the SNPs identified in PAP3 are located in 15% of the total coding region, towards the centre of the gene (*Figure 15*). However, the frameshift mutations we described are, as previously mentioned, not located in this region, but instead, unique to *PAP3,* are found in the R3 repeat region.

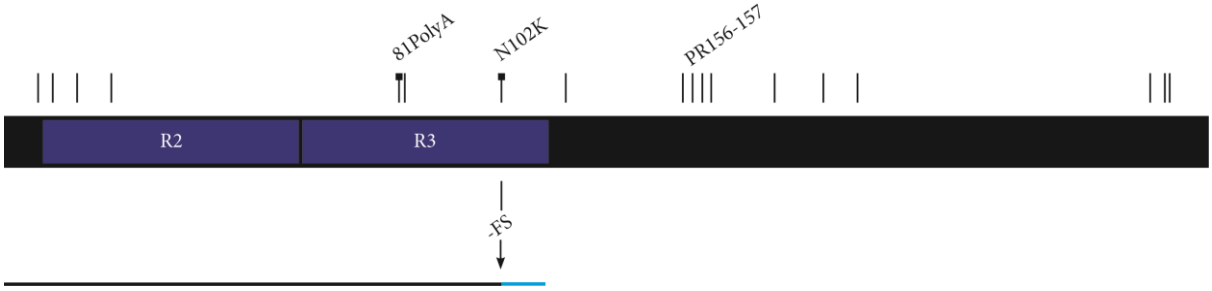
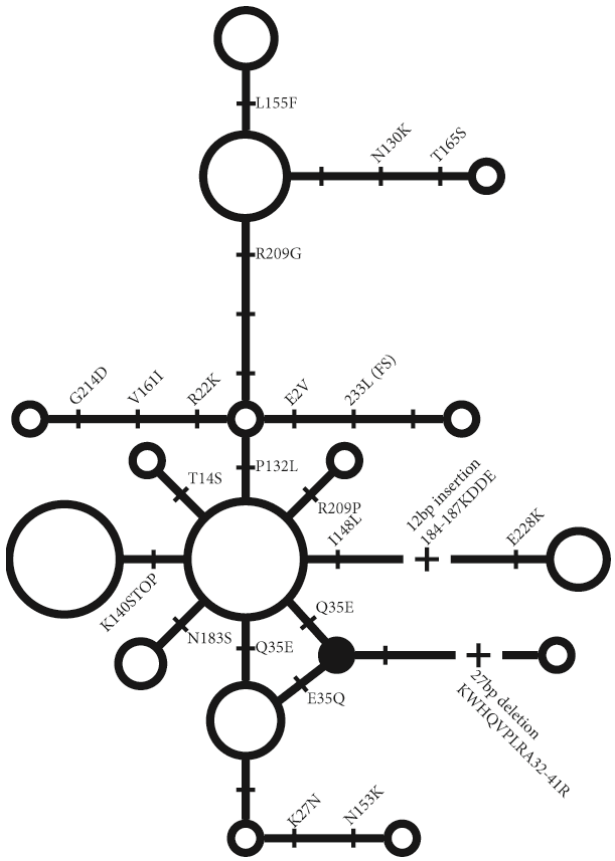


Figure 15 Schematic representation of the PAP3 protein showing positions of amino acid replacements and potentially functionally significant polymorphisms in 45 accessions. The horizontal black bar represents the full length protein, with the positions of the MYB domains (R2 and R3) shown as blue boxes. Each small vertical bar above the schematic indicates the position of an amino acid replacement table (Table 5). The bars with black boxes on top indicate alleles likely resulting in dead alleles. Below the schematic is shown the mutation likely resulting in a non-functional protein: '-FS' is the frameshift caused by a single bp deletion. The in-frame (black) and out-of-frame (blue) portions of the putatively truncated protein produced by the frameshift allele is shown as a horizontal line below the mutation. The 81PolyA insertion does not have the putative protein displayed as the nature of the mutation makes it difficult to determine the length of the putative protein.

3.3.2.4 PAP4. PAP4 appears to have experienced the least sequence conservation of the four *PAP* genes as the haplotype network generated for PAP4 is more fragmented than those of the other three (*Figure 16*). The network revealed 16 haplotypes, many of which are only of a few accessions. Gene-wide, nucleotide diversity of the inferred cDNA alignment of 47 accessions of PAP4 surveyed was 0.00390. The nucleotide diversity of synonymous sites was somewhat larger than that of nonsynonymous sites ($\pi(s) = 0.00559$, $\pi(a) = 0.00350$). Tajima's *D* was not significantly different from 0, though Fu and Li's *D** and *F** were. Again, we observed a high concentration of SNPs towards the centre of the coding region, downstream of the R2R3-MYB repeat regions; 48% of the nonsynonymous mutations are located in a region accounting for only 23% of the total coding sequence (*Figure 17*).



.

Figure 16 Median-joining haplotype network of *PAP3* coding region alleles. 16 haplotypes were identified based on inferred cDNA nucleotide sequence from 45 *Arabidopsis thaliana* natural accessions; haplotypes are represented by open circles, and their frequencies indicated by relative circle size. The black-filled circle represents a hypothetical, unsampled haplotype required to complete the network. Branch lengths and tick marks reflect the number of nucleotide changes separating haplotypes, with the exception of the dashed lines; these represent indels of varying lengths and are labelled accordingly.

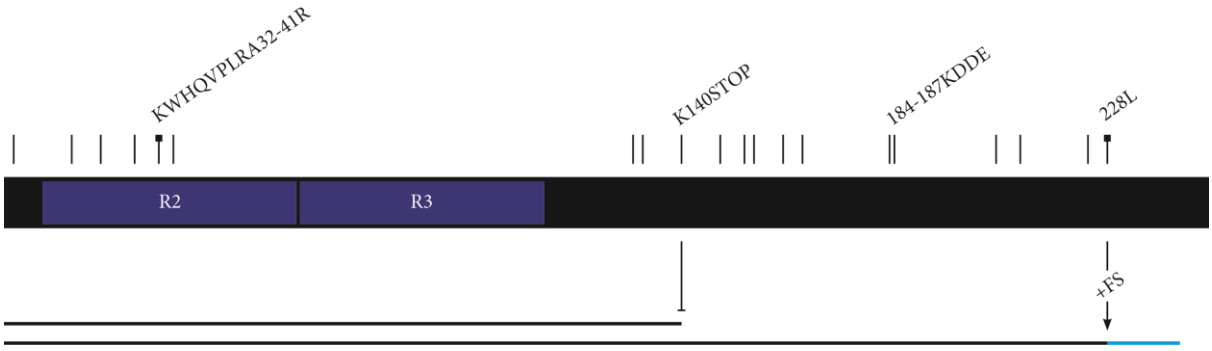


Figure 17 Schematic representation of the PAP4 protein showing positions of amino acid replacements and potentially functionally significant polymorphisms in 47 accessions. The horizontal black bar represents the full length protein, with the positions of the MYB domains (R2 and R3) shown as blue boxes. Each small vertical bar above the schematic indicates the position of an amino acid replacement table (Table 5). The bars with black boxes on top indicate alleles likely resulting in dead alleles. Below the schematic is shown the mutation likely resulting in a non-functional protein: '+FS' is the frameshift caused by an insertion. The in-frame (black) and out-of-frame (blue) portions of the putatively truncated protein produced by the frameshift allele is shown as a horizontal line below the mutation. The flat-bottomed bar below the schematic indicates the site of truncation of the protein in the ten accessions carrying the early stop codon.

The second most common haplotype, sampled from 10 accessions from our *PAP4* dataset, is defined by a mutation resulting in an early stop codon. The early stop codon reduces the length of the *PAP4* coding region in these accessions from 249 amino acids to 140 amino acids (*Table 5*). Notably, the only difference between this haplotype and the most common haplotype in the *PAP4* network is this early stop codon; no other polymorphisms were identified either upstream or downstream of the mutation. This mutation has previously been reported in *PAP4* of the Col-0 accession (Gonzalez *et al*., 2008); here, they noted that the *A. thaliana* accession Landsberg *erecta* (L*er*) is identical in sequence except for this early stop codon mutation. In our dataset, we observed 10 accessions with this mutation and none others, suggesting the mutation has occurred very recently.

We also observed a deletion beginning at amino acid site 32 in the R2 repeat region of accession N6. In examination of the cDNA sequence, it appeared the deletion was of 27 bp and inframe. However, examining the genomic alignment of our *PAP4* dataset revealed the deletion actually continued into intron 1 for another 17 bp. The new splice site for the boundary between exon 1 and intron 1 proposed by GENSCAN (Burge & Karlin, 1997) occurred 45 bp upstream of the splice site

found in all other accessions; this places the deletion in the centre of the R2 region of the MYB domain, and removes the second tryptophan residue necessary for proper folding and function from the protein sequence (Dubos *et al*., 2010). On its face, we would normally propose this to be a non-functional form of the *PAP4* gene. However, this deletion is the only mutation producing an alteration to amino acid sequence identified in N6, suggesting this mutation has occurred very recently. On the other hand, an in-frame insertion of 12 bp (184-187KDDE) was identified in three of the accessions (Can-0, Edi-0, Pi-0) from our dataset beginning at amino acid site 184. While these mutations weren't obvious candidates for being deleterious, these accessions share an additional two nonsynonymous SNPs, one upstream of the insertion, I148L, and one downstream, E228K. These two mutations fall well within the expected relative distance for compensatory mutations from deleterious mutations established by Davis *et al.* (2009).

We identified a frameshift mutation in a single accession, Ita-0, which is located towards the end of the coding region of PAP4, at amino acid site 233. A five base pair insertion resulting in a novel leucine residue (233L) also confers a premature stop codon at site 245, three residues short of the consensus stop. Still, it appears that sequence conservation has been maintained in this accession, which is not unexpected given the location of the insertion towards the end of the coding region. Other than the 233L insertion, we observed only one other unique mutation, the in-frame E2V SNP, and another SNP, P132L, shared by 13 other accessions in our dataset. To determine the influence the mutations of this particular accession had on our analysis of nucleotide diversity, we reanalysed π with Ita-0 excluded. Gene-wide, $\pi = 0.00382$, $\pi(s) = 0.00541$ and $\pi(a) = 0.00344$ and, perhaps most tellingly, Tajima's *D*, and Fu and Li's *D** and *F** were not significantly different from 0.

3.3.2.5 WER. The WER gene provides a stark contrast to the PAP gene family, with high conservation across the gene throughout the 48 accessions of our dataset. π of the WER coding region was 0.00040; $\pi(s) = 0.00136$ and $\pi(a) = 0.00017$. The WER network revealed only five haplotypes (*Figure 18*). 42 accessions in our WER dataset belong to the most common WER haplotype; the next most

common is represented by three accessions and the other three haplotypes have only one accession each belonging to them. The four minor haplotypes are differentiated from the most common haplotype by one SNP each, both of which occur outside of the MYB domain (*Figure 19*). Two haplotypes of one accession each, are differentiated by nonsynonymous SNPs, V5I in Can-0 and SG163R in Sav-0 (*Table 5*). The other two haplotypes are defined by synonymous mutations. Both of the nonsynonymous mutations are located in the 'undefined' region of the gene, outside of the R2R3-MYB regions.

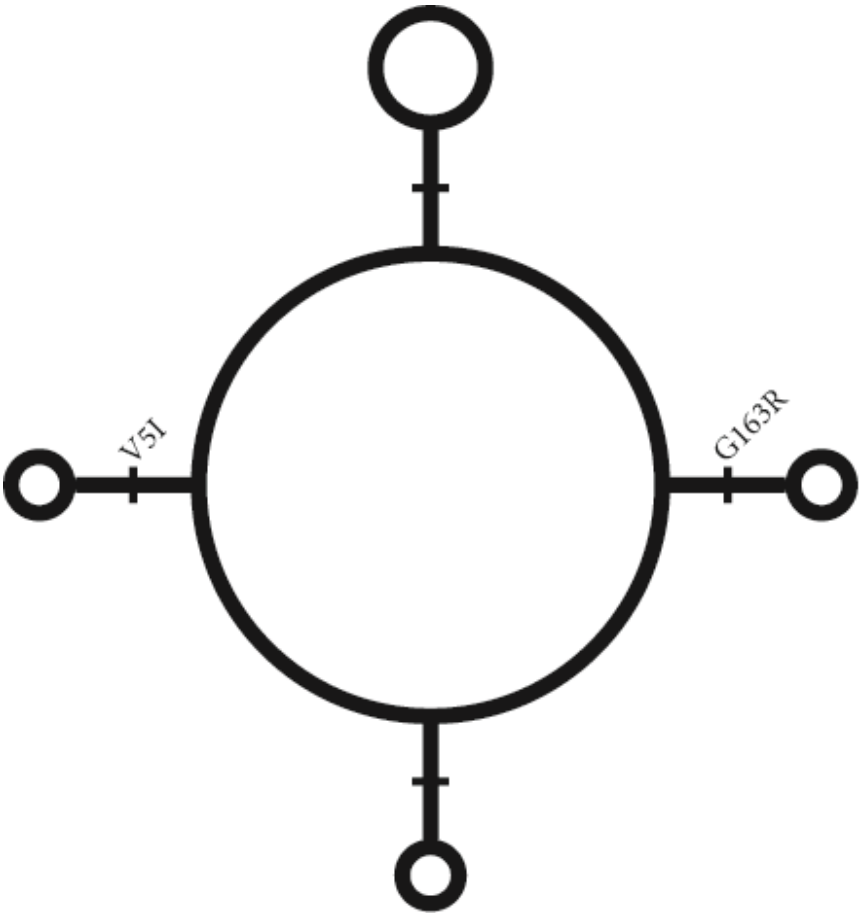


Figure 18 Median-joining haplotype network of *WER* coding region alleles. Five haplotypes were identified based on inferred cDNA nucleotide sequence from 48 *Arabidopsis thaliana* natural accessions; haplotypes are represented by open circles, and their frequencies indicated by relative circle size. Branch lengths and tick marks reflect the number of nucleotide changes separating haplotypes.

Figure 19 Schematic representation of the WER protein showing positions of amino acid replacements in 48 accessions. The horizontal black bar represents the full length protein, with the positions of the MYB domains (R2 and R3) shown as blue boxes. Each small vertical bar above the schematic indicates the position of an amino acid replacement table (Table 5).

3.3.2.6 R2R3 MYB regions. The PAP genes demonstrate high sequence conservation in the R2R3 repeat regions. All intragenic variation identified within the MYB regions was restricted to the R2 repeat region, except for PAP3 in accessions Sah-0 (a poly-A insertion beginning at amino acid site 81) and Rld-2 (an adenine insertion at amino acid site 116), both of which have been described in this work. We propose these mutations result in non-functional genes, as they both occur in the R3 repeat region of the PAP3 gene, vital for protein function; the poly-A drastically alters the protein sequence of the R3 region and the adenine insertion results in eight out-of-frame residues and an early stop codon on residue short of the end of the R3 region. Other than these two cases, no intragenic variation was found in the R3 region. This higher SNP incidence in the R2 region compared to the R3 region is the opposite result observed in the related gene R2R3-MYB gene *AtMYB0* (GL1), where the majority of R2R3 region mutations were found in the R3 repeat region (Bloomer *et al*., 2012).

The typical primary structure of the R2 repeat region is three regularly spaced tryptophan residues, with 19 amino acids dividing the first and second residues, and another 19 dividing the second and third (Dubos *et al*., 2010; Stracke *et al*., 2001). The R3 repeat region follows a similar pattern, though the first residue is typically replaced with a phenylalanine or isoleucine residue, and the required residues are separated by 18 amino acids each. These residues are the basis for the secondary helix-helix-turn-helix structure of the R2R3-MYB repeat regions. In our analyses, we noted PAP1, PAP2 and PAP3 had a leucine residue in the first required position, a favoured substitution for

isoleucine and unlikely to significantly impact protein structure or function. Both PAP4 and WER maintained the expected phenylalanine residue in the first required position.

3.3.2.6.1 PAP1. The forty-eight accessions of *A. thaliana* analysed show high sequence conservation in the R2 repeat region of PAP1. The only major deviation from total consensus is at amino acid site fifty-three, where 81% of accessions have a serine residue and the rest have an asparagine residue. The R2R3-MYB consensus sequence elucidated by Stracke *et al*. (2001) shows that the majority (54%) of R2R3-MYB genes in *A. thaliana* have an aspartic acid residue at this same site fifty-three, with the second most frequent residue (16%) asparagine. Serine, asparagine and aspartic acid are all polar/hydrophilic, so are favoured substitutes as they are expected to have little impact on protein structure and function (Betts & Russell, 2003). The only other deviation from intragenic consensus found was a histidine residue replacing a proline residue at site 30 in accession Cvi-0, a disfavoured substitution due to the large size and polar/hydrophilic properties of histidine compared to the small size and non-polar/hydrophobic properties of proline (Betts & Russell, 2003). This substitution occurs within the 'turn' region between the second and third tryptophan residues (sites 26 and 46, respectively) required for the helix-turn-helix secondary structure of the R2 repeat motif; this feature of the tertiary protein structure is known as the 'recognition helix' due to its direct role in identifying target DNA (Dubos *et al*., 2010). Given the location and type of mutation, there is high potential for disruption of protein folding likely negatively impacting protein function.

The R3 repeat region of PAP1 demonstrates complete intragenic consensus. However, there is a deviation from the expected amino acid sequence involved in protein folding compared to the Stracke *et al*. (2001) consensus sequence: in *A. thaliana,* the three tryptophan residues required for the helix-helix-turn-helix secondary structure are conserved in the R2 repeat regions of all R2R3-MYB genes so far analysed (Stracke *et al*., 2001). In the R3 repeat regions of the R2R3-MYB genes analysed by Stracke *et al*. (2001), the first tryptophan residue is replaced with a phenylanine residue in 69% of *A. thaliana* R2R3-MYB cases and an isoleucine residue in 14% of cases. This is seemingly

unique to the R3 region of R2R3-MYB genes in *A. thaliana,* as the three tryptophan residues have been found to be conserved in R1, R2 and R3 repeat regions in non-R2R3-MYB *A. thaliana* genes, as well as in other species, including *Drosophila,* mouse and human (Ogata *et al*., 1996). In the PAP genes, we observed PAP1*,* PAP2 and PAP3 all had a leucine residue at this site, whereas PAP4 had a phenylalanine residue, predicted by the Stracke *et al.* (2001) consensus. All four amino acids found at this site are hydrophobic, though phenylalanine and tryptophan are aromatic, whereas isoleucine and leucine are aliphatic.

3.3.2.6.2 PAP2. Only one site, amino acid sixteen, is variable in the R2 repeat region of PAP2; where 96% of the accessions examined had a leucine residue, 4% had a glutamine residue. While the leucine residue is clearly in the majority in PAP2, all accessions of PAP1*,* PAP3 and PAP4 have a glutamine residue at their corresponding amino acid site in the R2 repeat region. The majority residues at this same site in *A. thaliana* R2R3-MYB genes overall are asparagine and serine (23% and 20%, respectively), both of which are hydrophilic; the majority residue in the PAP genes, glutamine, is also hydrophilic, whereas leucine is hydrophobic (Stracke *et al*., 2001). Based on this, it would appear that the ancestral residue in the PAPs is the hydrophilic residue glutamine, but in PAP2 the hydrophobic leucine residue appears to be fixed at this site.

There is complete intragenic consensus across the R3 repeat region in PAP2. Of the three expected tryptophan residues involved in the helix-turn-helix, the latter two are present, with the first being replaced by a leucine residue, also seen in PAP1 and PAP3. Again, this is not entirely unexpected, given the similarity in properties of leucine to isoleucine, phenyalanine and tryptophan, which are the expected residues at this site (Stracke *et al*., 2001). In comparison to the other PAP genes, we didn't observe changes we expect would cause major disruption of structure or function: in all PAP2 accessions we found the hydrophilic arginine residue at site five, whereas at the same site in the accessions of the other three PAP genes lysine is present, which is also hydrophilic; at amino

acid site eight in PAP2 we found a hydrophobic asparagine residue in all accessions, whereas the similarly hydrophobic serine residue is found at the same site in all.

3.3.2.6.3 PAP3. We observed three sites in the R2 repeat region of PAP3 which display intragenic variability. The first, K10R, was identified in three accessions. In all accessions of PAP1, PAP2, PAP4 and the other 42 accessions of the 45 analysed of PAP3, a lysine residue is found at site 10, making the arginine residues novel, though likely of low impact to protein structure and function. Notably, Rld-2, which was excluded from analysis in this case, also had the novel K10R mutation, again hinting at the more recent timing of the N102K frameshift mutation.

The second, and most frequent mutation in the R2 region of PAP3, is T15A, identified in 15 accessions. Interestingly, the other 30 accessions maintain a threonine residue at this site, as do all PAP1 alleles analysed here. Conversely, all accessions analysed of PAP2 and PAP4 have an alanine at the same site. The majority residue at site 15 in PAP3, threonine, is in the minority when compared to the other three PAP genes; 40% of analysed *A. thaliana* accessions have the threonine allele, while 59% have the alanine allele. According to Stracke *et al*. (2001), the most common residue at this site is across *A. thaliana* R2R3-MYB genes is proline (44%), though alanine, the PAP majority allele, is represented in 14% of cases. While threonine was not represented in the consensus, it is not an unlikely candidate given the similar properties of the three residues. Further, we observed a serine residue, which is again a similar residue to the aforementioned three, at the same site of PAP3 in one accession, Can-0, an island population. A third SNP identified in the R2 region shares similarities to the high frequency T15A SNP in that it is a replacement of an amino acid with another of similar properties, and therefore unlikely to be of dramatic consequence. The SNP, R22S, is found in one accession of PAP3, Akita.

PAP3 is unique as it is the only PAP gene where mutations were found in the R3 repeat region of the accessions in our dataset. Two cases were observed, 81poly-A (and as a direct result of this, G82R) in Sah-0, and N102K in Rld-2. In both cases, we deem it highly likely the genes are rendered

non-functional, due both to the location of the mutation in crucial regions of the gene and the shift in reading frame. Other than these two cases, the R3 repeat region of PAP3 exhibits 100% intragenic conservation as is observed in the other three PAP genes.

3.3.2.6.4 PAP4. PAP4 claims both the most SNPs over all and in the R2 repeat region. Still, only one of the five mutations identified in this region has a high frequency. The SNP Q35E affects six accessions. Four of the five mutations identified in the R2 region of PAP4 are likely low impact, as the replacement residues share properties with the residues they replace. However, we did observe a mutation in the R2 region unique to PAP4: KWHQVPLRA32-41R is the only indel in the R2R3-MYB repeat regions of any accession of the four PAP genes analysed; still, as previously mentioned, this may not be the final amino acid sequence, as splice site predictions put the end of exon 1 upstream of this site. The deletion is the only mutation observed in accession N6 across the entire *PAP4* gene, indicating the mutation is either also a low impact mutation or is relatively recent and further mutations have not yet accumulated. The latter is the most likely scenario, as the mutation deletes the second of the three necessary tryptophan residues for helix formation in the R2 region, highly likely to have a deleterious impact on protein structure and function (Dubos *et al*., 2010).

3.3.2.6.5 WER. We observed complete intragenic consensus of the R2R3-MYB regions of WER across the 48 *A. thaliana* accessions analysed. Comparing the R2 and R3 repeat regions of WER to the PAP genes, we noted that the R3 repeat region shared 76% residue identity to the consensus sequence between the PAP genes, whereas the R2 repeat region only shared 50% residue identity, likely reflecting the capacity to associate with the same bHLH proteins while targeting different regulatory sequences. In the R2 repeat region, 60% of the shared residues are downstream of the second required tryptophan reside, particularly around the third tryptophan, in the turn region of the helix-turn-helix structure, likely reflecting the imperative of maintaining this region for protein function.

Regarding specific DNA recognition by the R2R3-MYB region, it has been previously elucidated in humans, mice and *Drosophila* MYB regions that while both the R2 and R3 repeat regions are structurally similar and both involved in DNA recognition, they differ in thermal stability; typically, the three-dimensional structure of R2 forms a hydrophobic cavity which confers thermal instability in the DNA-free state, suggesting increased conformational flexibility (Ogata *et al.*, 1996). Of the ten residues cited as necessary for hydrophobic core formation, six are conserved in the R2R3-MYB PAP genes we analysed. Typically, as in humans, mice and *Drosophila,* a valine residue at amino acid site 14 is conserved, responsible for cavity formation; Ogata *et al*. (1996) noted that replacement of this valine residue with a leucine residue is sufficient to fill the cavity in R2, resulting in thermal stability. We observed a leucine residue at site 14 in all accessions of the four PAP genes, indicating low conformational flexibility in the R2 regions of the PAP genes due to thermostability.

3.3.3 Intergenic molecular evolution amongst the PAPs

To evaluate regions patterns of purifying selection and deviation from such, we aligned the PAP genes in pairs and performed sliding window analyses of *Ka/Ks*. A common pattern of selection emerged, where the R2R3-MYB regions revealed evidence for purifying selection (*Ka/Ks* « 1), punctuated by peaks of departure from purifying selection typically at the beginning of the R2 and R3 regions of the MYB domain. Immediately downstream of the conserved MYB domain, a region of apparent positive selection was observed in all comparisons; for the remainder of the 'undefined' region, selection appeared to fluctuate randomly and overall *Ka/Ks* ≈ 1. Comparing the *PAP* genes to *WER* revealed a similar pattern to that of the *PAP* gene comparisons, where we saw negative selective pressures in the R2R3 MYB region and release of selective constraint in the undefined domains (*Figure 2.18*). Deviation of *Ka/Ks* from 0 is predominantly affected by the *PAP* genes, as the accessions in our dataset demonstrate near complete sequence conservation at the *WER* locus.

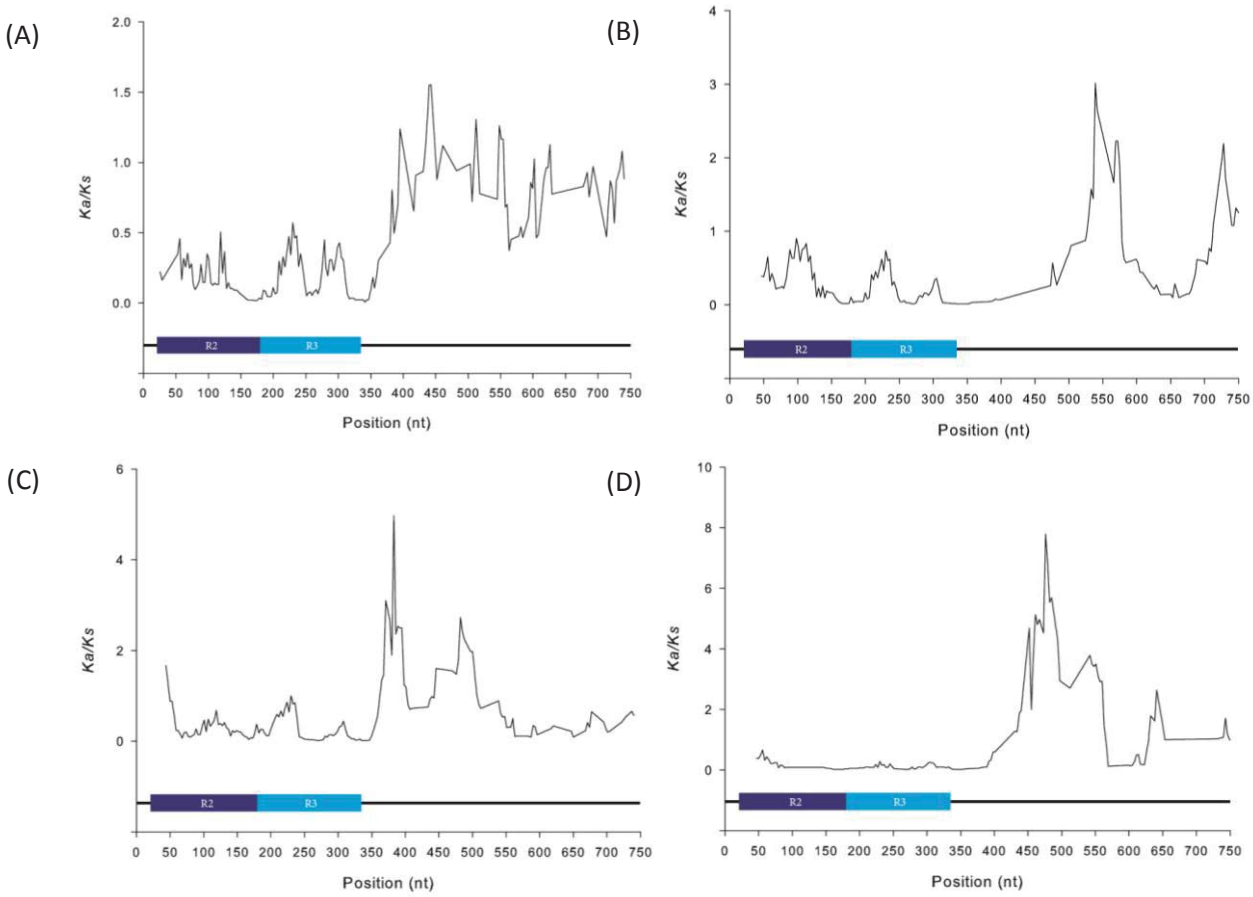


Figure 20 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of (A) *PAP1* and *WER*, (B) *PAP2* and *WER,* (C) *PAP3* and *WER* and (D) *PAP4* and *WER* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.1 PAP1 and PAP2. We saw the expected pattern of apparent negative selection in the R2R3 MYB domain followed by release of selective constraint outside of this region in the 'undefined' domain (*Figure 21*). The average *Ka/Ks* value for the R2R3 MYB domain was 0.15 whereas the average for the 'undefined' domain was 0.73 (values of two major peaks were removed). A summary of *Ka/Ks* values for *PAP* gene comparisons can be found in Table 8. Over the whole alignment we observed two major peaks of positive selection, the first (at its highest, *Ka/Ks* = 22.95) covers six overlapping windows from 157-211 bp; underlying this peak is a high frequency polymorphism of *PAP1* at the last residue of the R2 region in nine accessions, S60N, a polymorphism which is partially responsible for the differentiation of the two *PAP1* haplogroups. Where the majority allele of both

PAP1 and *PAP2* is a serine residue, the nine aforementioned accessions of *PAP1* have acquired an asparagine residue at this site. The second peak (*Ka/Ks* = 59.25) covers a single window, from 403-442 bp. Again, underlying this peak is a polymorphism of *PAP1* in nine accessions; where *PAP2* has an isoleucine residue and the majority allele of *PAP1* is an asparagine residue at this site, the same nine accessions underlying the previous peak have a threonine residue instead. Immediately outside the MYB domain is a region of moderate positive selection, from *Ka/Ks* = 1.09 to *Ka/Ks* = 3.71, covering 19 overlapping windows from 358-451 bp; the largest peak in this alignment (*Ka/Ks* = 59.25) also falls in this region. It appears that underlying this site of positive selection are several differences in sequence between *PAP1* and *PAP2*, as there are no non-synonymous intergenic differences in this region, except for the single high frequency polymorphism towards the end of this area, already previously accounted for, and serine residue at amino acid site 171 in *PAP2* where there is no corresponding residue in *PAP1*. While there several other high frequency polymorphic sites in *PAP1* contributing to the division between P1A and P1B, these are apparently of less impact than the previously mentioned cases. For example, a moderate peak of positive selection covering four overlapping windows from 640-688 bp is accounted for by two high frequency polymorphisms of the same ilk as the two sites underlying the two major peaks of this alignment. The polymorphisms, a methionine residue where the majority allele of *PAP1* is an isoleucine residue and *PAP2* has an alanine residue; and a threonine where the majority allele of *PAP1* and *PAP2* have an alanine residue, are apparently of low impact as, at its highest, *Ka/Ks* = 2.1 in this area. Other than these few sites, the mild variation in *Ka/Ks* values can be attributed to sequence variation between the two genes.

Table 7 Summary of *Ka/Ks* averages of different gene regions of the *PAP* genes

Comparison	Average Ka/Ks(MYB)	Average Ka/Ks(Undefined)
PAP1:PAP2	0.15	0.73
PAP1:PAP3	0.13	0.6
PAP1:PAP4	0.09	0.71
PAP2:PAP3	0.2	0.95
PAP2:PAP4	0.19	0.78
PAP3:PAP4	0.09	0.79

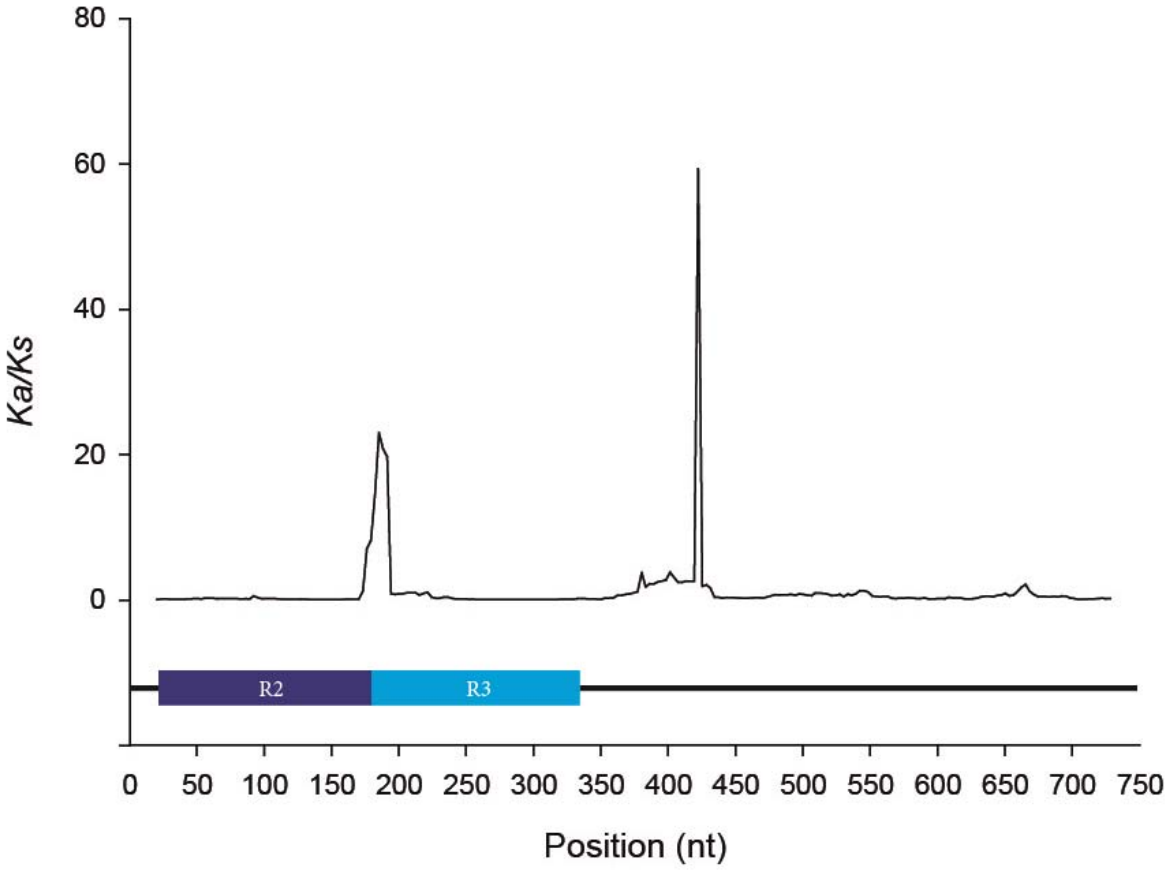


Figure 21 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP1* and *PAP2* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.2 PAP1 and PAP3. The pattern of negative selection pressure in the R2R3 MYB domain compared to relaxed selective pressure in the 'undefined' domain was easier to visualise in the comparison of *PAP1* and *PAP3* as the peaks of positive selection were much reduced for this alignment compared to that of *PAP1* and *PAP2* (*Figure 22*). The MYB domain on the whole demonstrated negative selective pressure (average *Ka/Ks* = 0.13) compared to the 'undefined' domain (average *Ka/Ks* = 0.6). The pattern of low *Ka/Ks* values in this area was perturbed early in the R2 region by a high frequency polymorphism in *PAP3* affecting 16 accessions, where 15 have an alanine residue and one a serine residue rather than a threonine residue which is found in the remaining *PAP3* accessions and all *PAP1* accessions, as well as several singleton polymorphism found

in *PAP1* and *PAP3*. A second mild perturbation in this region occurs as a result of a high frequency polymorphism already seen in the *PAP1* and *PAP2* comparison, affecting nine accessions at the last residue of the R2 region. As seen in the *PAP1* and *PAP2* alignment, the highest peak was found in the region of relaxed selection immediately outside of the R2R3 MYB domain; at its highest, *Ka/Ks* = 3.19. The likely underlying cause of this peak is the deletion of six nucleotides from *PAP3* in four accessions. Underlying a second peak (*Ka/Ks* = 2.76) is a high frequency polymorphism of *PAP1*, where nine accessions of *PAP1* a serine residue where the rest of the *PAP1* accessions and all *PAP3* accessions have a threonine residue. Even aside from these cases, this area of moderate positive selection at the beginning of the 'undefined' domain, spanning 10 overlapping windows, from 382-457 bp, contains multiple polymorphisms and mutations, such as a histidine residue at the first site outside of the MYB domain in *PAP3* where *PAP1* has no residue, a second high frequency polymorphic site of *PAP1* and a number of singleton polymorphisms in both *PAP1* and *PAP3*. We observed a curious region of apparent negative selection in the 'undefined' domain, spanning 15 overlapping windows, from 535-622 bp. This may be somewhat artificial as, although there appears to be a greater number of synonymous changes to synonymous sites compared to the surrounding regions, there are also a number of non-synonymous changes to non-synonymous sites. However, as reflected by the *Ka/Ks* values of the area, the former exceed the latter. In this case, it seems that it is a region subject to a high rate of mutation, though the incidence of synonymous change has exceeded non-synonymous by chance.

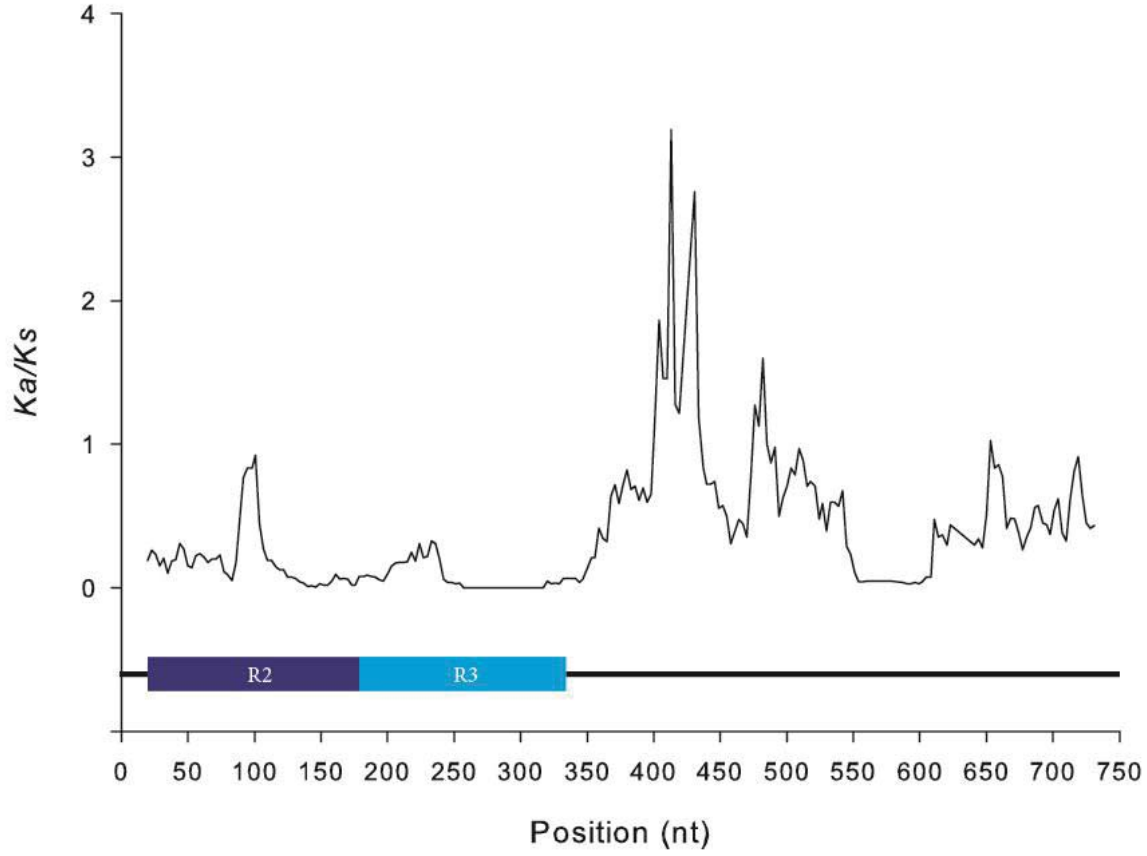


Figure 22 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP1* and *PAP3* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.3 PAP1 and PAP4. As with the comparison between *PAP1* and *PAP3*, we clearly saw the dichotomy between the R2R3 MYB domains and the 'undefined' domains in the comparison of *PAP1* and *PAP4* (*Figure 23*). The only noticeable deviation from strong negative selective pressure in the R2R3 domain is found early in the R2 region, the result of a high frequency polymorphism in *PAP4* in 6 accessions, a 27 bp deletion from *PAP4* in a single accession and a singleton polymorphism in *PAP1*. Another deviation from the standard negative selection (*Ka/Ks* < 0.3) is the result of a previously observed high frequency polymorphism of *PAP1*. Through the entire R2R3 MYB domain, *Ka/Ks* failed to exceed 0.5. So far, this comparison demonstrated the greatest contrast between the MYB domain, where average *Ka/Ks* = 0.09, and the 'undefined' domain, where average *Ka/Ks* = 0.71,

as well as the greatest negative selection pressure in the MYB domain. Again, immediately outside the MYB domain, we observed a region of relaxed selective constraint, covering 26 overlapping windows from 352-466 bp, wherein the two highest peaks of the alignment were observed (*Ka/Ks* = 2.97, 2.93). Underlying these peaks are two high frequency polymorphisms in *PAP4* and one in *PAP1*. The first polymorphism of *PAP4*, affecting 12 accessions, is a leucine residue instead of a proline residue in *PAP4* for the majority of accessions in this dataset and *PAP1* for all accessions; the second of *PAP4*, affecting 10 accessions, is an early stop codon in place of a lysine residue for the majority of accessions in this dataset and *PAP1* for all accessions. The high frequency polymorphism of *PAP1* affects nine accessions which have a threonine residue in place of an asparagine residue for *PAP1* in the majority of accessions. A wide peak of apparent neutral selection (at its highest, *Ka/Ks* = 1.28) covering nine overlapping windows from 490-577 bp, corresponds to a region of noticeable variation, wherein we observed several singleton polymorphisms in both *PAP1* and *PAP4*, a low frequency polymorphism in *PAP4* (affecting three accessions), a high frequency polymorphism in *PAP1* (affecting nine accessions) and a 12 bp insertion mutation producing amino acid sequence KDDE in *PAP4* of three accessions, which appears to be a repetition of the sequence adjacent to it. In this comparison between *PAP1* and *PAP4*, we noted with interest that although this insertion in *PAP4* appears novel when considered alone, *PAP1* in all accessions from our dataset maintain the first half of this insertion, KD, without immediate repetition. Further downstream, another peak of positive selection (*Ka/Ks* = 1.73) was observed due to an underlying high frequency polymorphism in *PAP4*, affecting 10 accessions, nine of which have glycine residue and one which has a proline residue in place of the majority residue arginine.

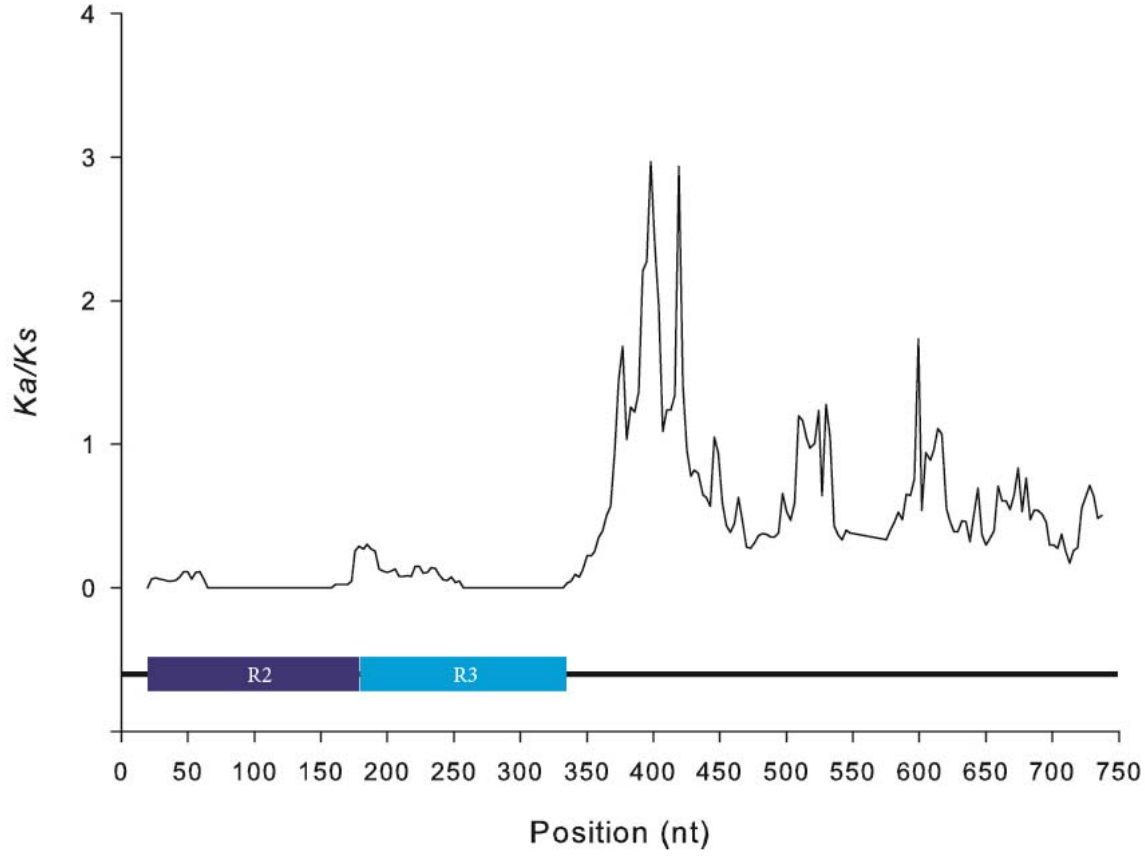


Figure 23 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP1* and *PAP4* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.4 PAP2 and PAP3. Maintaining the expected pattern of strict negative selection in the MYB domain (average *Ka/Ks* = 0.20, peaks removed) contrasting with the relaxed selective pressure in the 'undefined' domain (average *Ka/Ks* = 0.95, peaks removed), we observed two peaks of positive selection in the R2 region of the MYB domain which dwarf all others (*Figure 24*). The first peak (*Ka/Ks* = 11.52), covering two overlapping windows from 28-70 bp, is the result of a low frequency polymorphism of *PAP3*, affecting three accessions and producing an arginine residue instead of a lysine residue, and a high frequency polymorphism affecting of PAP3, affecting 16 accessions, 15 of which maintain an alanine residue and one which has a serine residue in place of a threonine residue. *PAP2* has an alanine residue, the minority allele for *PAP3*, at this site in all accessions of our

dataset. The second peak (*Ka/Ks* = 13.07), covering a single window from 73-112, corresponds to a low frequency polymorphism in *PAP2* affecting two accessions, which maintain a glutamine residue in place of the leucine residue found at this site in all other accessions in our dataset; at the corresponding site of *PAP3*, all accessions from our dataset had a glutamine residue, the minority allele in *PAP2,* similar to what was seen in the previous incidence of positive selection. Both of the accessions affected this polymorphism have a second polymorphism 6 bp downstream which produces a gycine residue instead of an aspartic acid residue. We also observed a singleton polymorphism one amino acid upstream of the corresponding site in *PAP3*, resulting in a serine in place of an arginine residue. We noted a second area of the MYB domain which fluctuated from the expected tight negative selective pressure (0.8 > *Ka/Ks* > 0.5), covering six overlapping windows from 184-238 bp. Corresponding to the beginning of the R3 region of the MYB domain, this area has a number of intergenic amino acid differences whereas the remainder of the R3 region (111 bp of 153 bp total) has a number of silent differences though no amino acid differences. Again, we observed a region of relaxed selective pressure outside the MYB domain, covering 27 overlapping windows from 361-487 bp (at its peak, *Ka/Ks* = 2.40), the result of a number of intergenic amino acid differences, several discrete polymorphisms and a low frequency deletion of 6 bp in *PAP3*, affecting four accessions. Underlying another peak of positive selection (*Ka/Ks* = 3.08), covering three overlapping windows from 565-616, is a high frequency polymorphism of *PAP2*, where nine accessions have a glycine residue instead of a glutamic acid residue seen in other accessions at this site in *PAP2* and the corresponding site in *PAP3* for all accessions in our dataset. Adjacent to this site is an insertion of 12 bp in *PAP2* relative to *PAP3*. The final peak, largest in the 'undefined' domain (*Ka/Ks* = 4.2), covering two overlapping windows from 703-745 bp, is the result of two adjacent polymorphisms of *PAP3*, the first affecting two accessions (a leucine residue in place of a phenylalanine residue in *PAP3* for the remaining accessions and *PAP2* for all accessions in our dataset) and the second affecting 11 accessions (a glutamic acid residue in place of an aspartic acid residue in *PAP3* for the remaining accessions and *PAP2* for all accessions in our dataset).

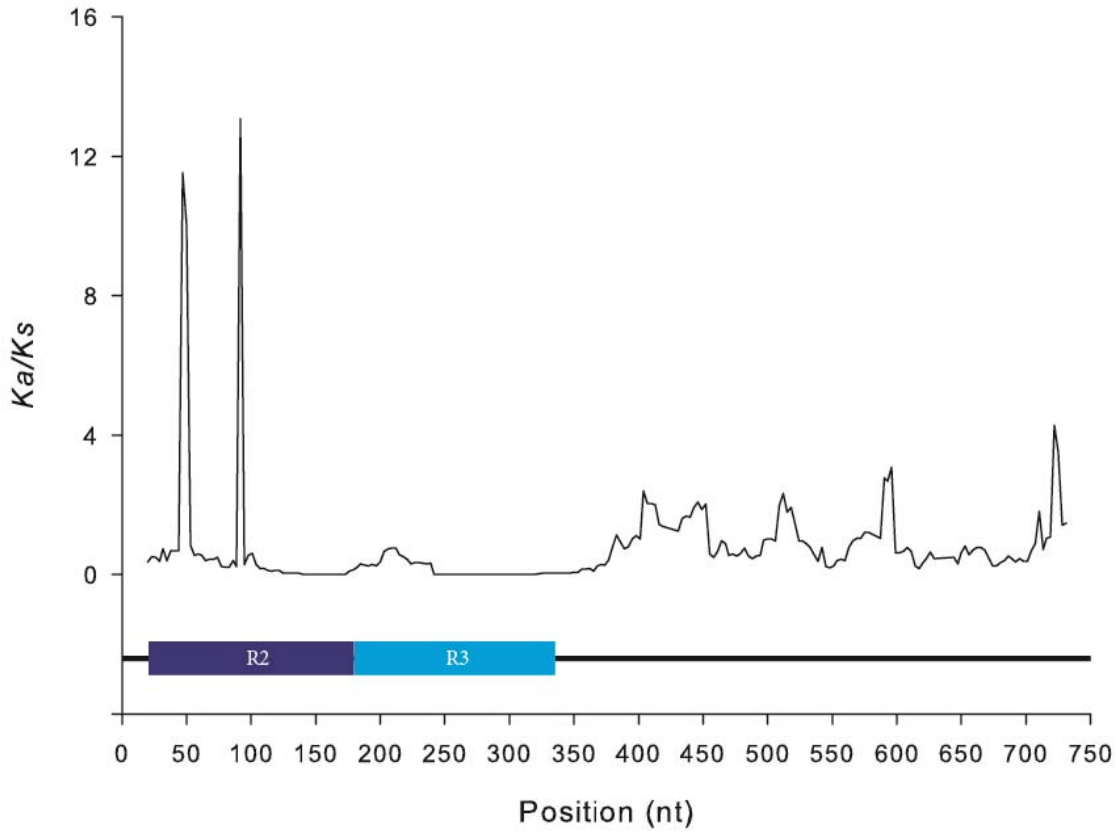


Figure 24 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP2* and *PAP3* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.5 PAP2 and PAP4. Here we again observed a fluctuation away from the overall expected negative selective pressure in the MYB domain, particularly at the beginning of the R2 and R3 regions (*Figure 25*). The first area of relaxed selective pressure (*Ka/Ks* = 0.59), covering eight overlapping windows from 40-130 bp, covers an area with a number of discrete polymorphisms in both *PAP2* and *PAP4*, a high frequency polymorphism in *PAP4* and a deletion of 27 bp from the *PAP4* gene in one accession. As seen previously, underlying the peak of apparent neutral evolution at the beginning of the R3 region (*Ka/Ks* = 1.193), covering nine overlapping windows from 163-226 bp, are a number of intergenic amino acid differences coupled with complete intragenic consensus. The remainder of the R3 region (129 bp of 153 bp total) shows some silent mutation while the amino

acid sequences of both genes are identical. The region of relaxed selection pressure immediately outside the MYB domain, covering 18 overlapping windows from 361-451 bp, is location of the highest peak of positive selection of this alignment (*Ka/Ks* = 5.52), covering three overlapping windows from 382-427 bp. Underlying this peak are two high frequency polymorphisms in *PAP4*. The first, affecting 12 accessions, maintains a leucine residue where a proline residue is found in other accessions and at the corresponding site of *PAP2* for all accessions in our dataset. The second results in an early stop codon for the 10 affected accessions, where the majority allele for *PAP4*, and *PAP2* of all accessions in our dataset, is a lysine residue. Another peak towards the 3' end of the coding sequence (*Ka/Ks* = 2.51), covering 11 overlapping windows from 652-721, appears affected by a number of intergenic amino acid changes as well as a low frequency polymorphism in *PAP4*, affecting 3 accessions, where a lysine residue has replaced the majority allele residue glutamic acid in *PAP4*; all accessions of our dataset also had a glutamic acid at the corresponding site in *PAP2*. The average differences in overall selective pressures for the two domains of the coding region are similar to previously seen, with *Ka/Ks*(MYB domain) = 0.19 and *Ka/Ks*('undefined' domain) = 0.78 (peaks removed).

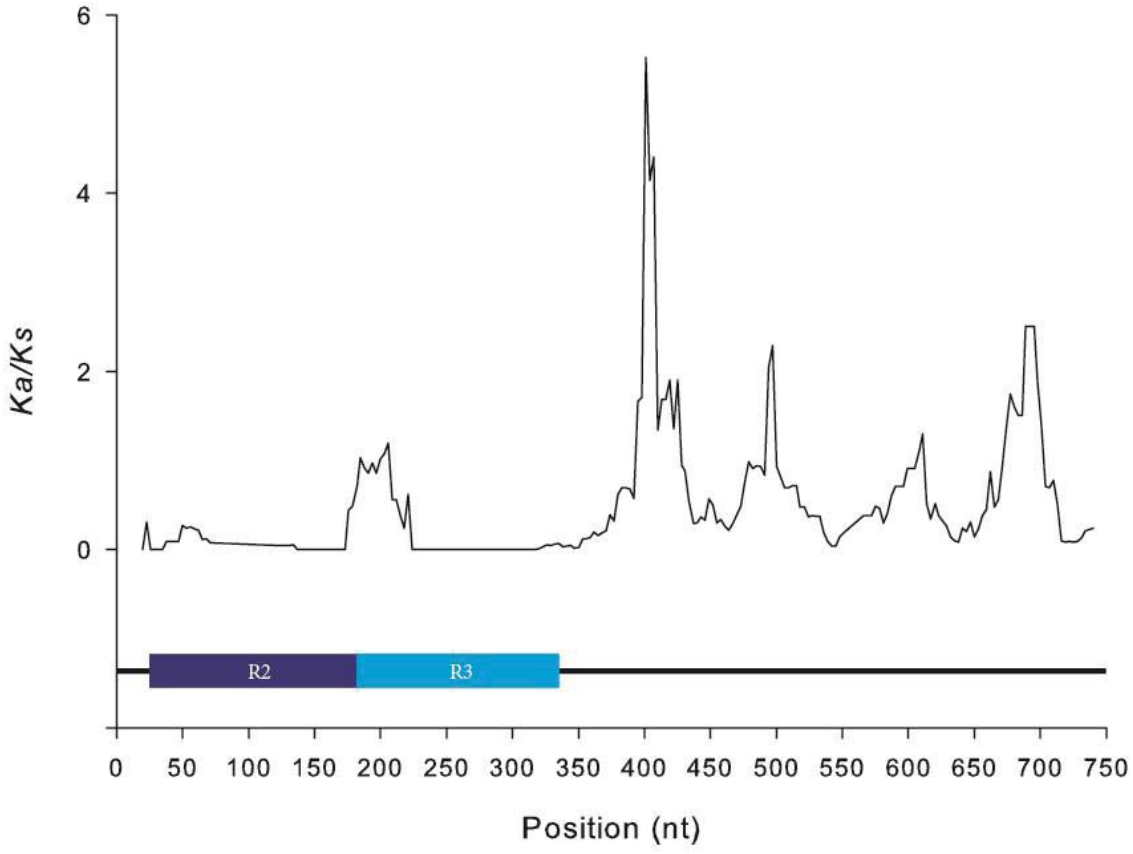


Figure 25 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP2* and *PAP4* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.3.6 PAP3 and PAP4. We again see the typical pattern of negative selection in the MYB domain (average *Ka/Ks* = 0.09, peaks removed) contrasting with the relaxed selection observed in the 'undefined' domain (average *Ka/Ks* = 0.79, peaks removed) (*Figure 26*). Punctuating this expectation are three major peaks of positive selection, one at the beginning of the R2 region and two in the 'undefined' domain (*Ka/Ks* = 17.81, 12.38). The first peak (*Ka/Ks* = 11.44), covering a single window from 28-67 bp, is the result of a high frequency polymorphism is *PAP3*, affecting 16 accessions, 15 of which have an alanine residue and one which has a serine residue where the rest of the accessions have a threonine residue at this site in *PAP3*; all accessions have an alanine residue at the corresponding site in *PAP4*. Surrounding this peak, an area covering 12 overlapping windows from 1-

73 bp, a low frequency polymorphism in *PAP3*, resulting in an arginine residue in the place of a lysine residue in the three accessions affected, as well as three discrete polymorphisms in *PAP4*, result in a deviation from negative selection. Other than this, *Ka/Ks* is less than 2.5 for the remainder of the domain, the lowest consistent value for the MYB domain of all comparisons. The second peak (*Ka/Ks* = 17.81), covering four overlapping windows from 385-442 bp, is located in the region of relaxed selective pressure immediately outside the MYB domain, covering 17 overlapping windows from 358-454 bp. Underlying this peak are several polymorphisms in both *PAP3* and *PAP4*. A 6bp deletion, affecting four accessions, and a SNP resulting in the replacement of an arginine residue found at this site in both *PAP3* and *PAP3* with a lysine residue, are found in *PAP3*; the replacement of a proline residue with a leucine residue, affecting 12 accessions, and the replacement of a lysine residue with an early stop codon, affecting 10 accessions, is also found in this area in *PAP4*. Aside from these polymorphisms, accounting for the peak of positive selection, an aspartic acid in *PAP3* where there is no corresponding amino acid in *PAP4*, and discrete polymorphism as well as a number of intergenic amino acid changes contribute to the increased *Ka/Ks* values in this region. Underlying the third major peak (*Ka/Ks* = 12.38), covering three overlapping windows from 601-661 bp, is a single high frequency polymorphism in *PAP4*, affecting 11 accessions, 10 of which have a glycine residue and one which has a proline residue whereas the other accessions in our dataset have an arginine residue at this site in PAP4. Towards the 3' end of the coding region, we also observed an area of neutral and positive selection (2.45 > *Ka/Ks* > 1.15), covering nine overlapping windows from 664-727 bp, the result of three polymorphisms. One, in *PAP3*, affects two accessions; the phenylalanine residue found at this site in the majority of cases in *PAP3* and all analysed cases in *PAP4*, has been replaced by a leucine residue. Adjacent to this, both *PAP3* and *PAP4* maintain an aspartic acid residue, though 11 accessions have a glutamic acid residue at this site in *PAP3*. The other, in *PAP4*, is the replacement of glutamic acid with a lysine residue in three accessions.

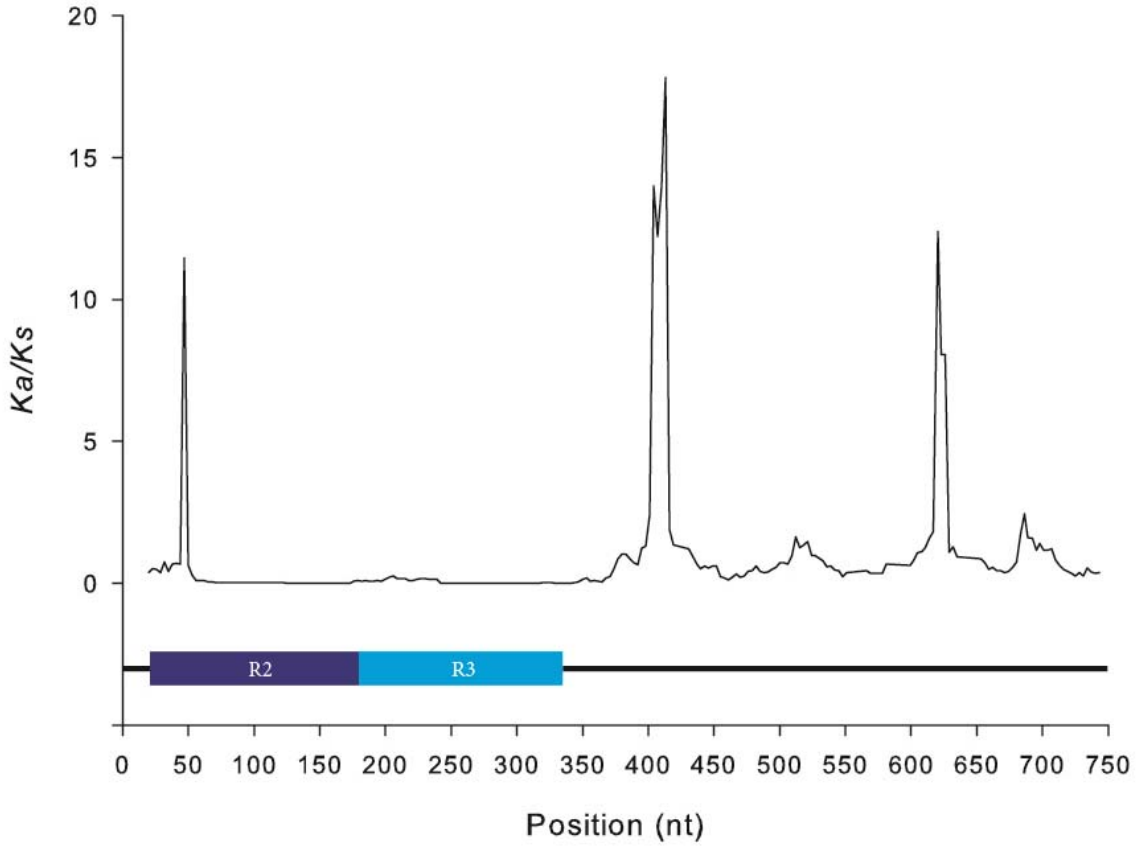


Figure 26 Sliding window analysis of *Ka/Ks* between inferred coding sequence alignments of *PAP3* and *PAP4* with *Ka/Ks* plotted against window midpoint. The underlying schematic indicates the inferred coding region (black lines) of the consensus sequences with the positions of the R2R3 MYB domains indicated by the overlaid boxes.

3.3.4 Analysing the phylogenetic relationships of the PAP genes

We investigated the phylogenetic relationships between the *PAP* genes in order to infer the duplication history of the *PAP* gene family. Bayesian analysis of the inferred coding sequences of the four *PAP* genes, using *AtMYB82* as an outgroup, identified a common ancestor for the four *PAP* genes but no further resolution of the relationships amongst the four, resulting in a four-way polytomy; based on branch length alone, however, it could be inferred that *PAP3* was the most diverged of the four duplicates, *PAP1* being next, and *PAP2* and *PAP4* being the least diverged (*Figure 27*). Still, because we observed levels of intergenic sequence variation expected to be sufficient for inference of phylogeny, while also observing intragenic regional variation in the

likelihood of polymorphisms occurring, we hypothesised that different regions within the genes were conflicting, disallowing accurate resolution of their phylogenetic relationships and returning the observed polytomy. To test this, we performed a Bayesian analysis of the R2R3-MYB regions only of the four *PAP* genes, *AtMYB82* as an outgroup. The resulting tree supported our supposition on the *PAP* gene family evolutionary history with increased resolution of the relationships, proposing a common ancestor of *PAP2* and *PAP4*, grouping *PAP1*, *PAP2* and *PAP4* together with the exclusion of all others, and grouping all the PAPs together with the exclusion of *AtMYB82* (*Figure 28*).

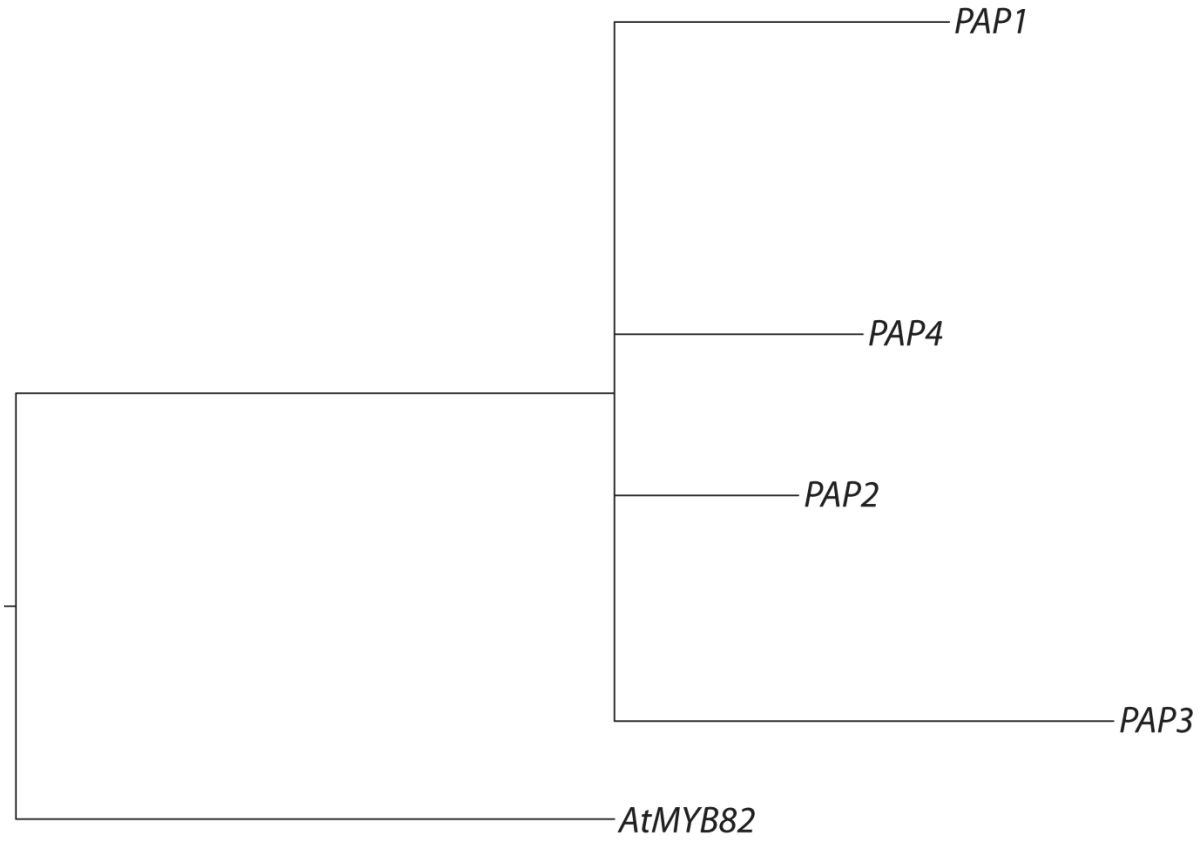


Figure 27 Bayesian phylogeny of consensus sequences of genomic alignments of the *PAP* genes. *AtMYB82,* the most closely related *Arabidopsis thaliana* MYB gene to the *PAP* genes, was used as an outgroup. The gene sequence was taken from www.arabidopsis.org.

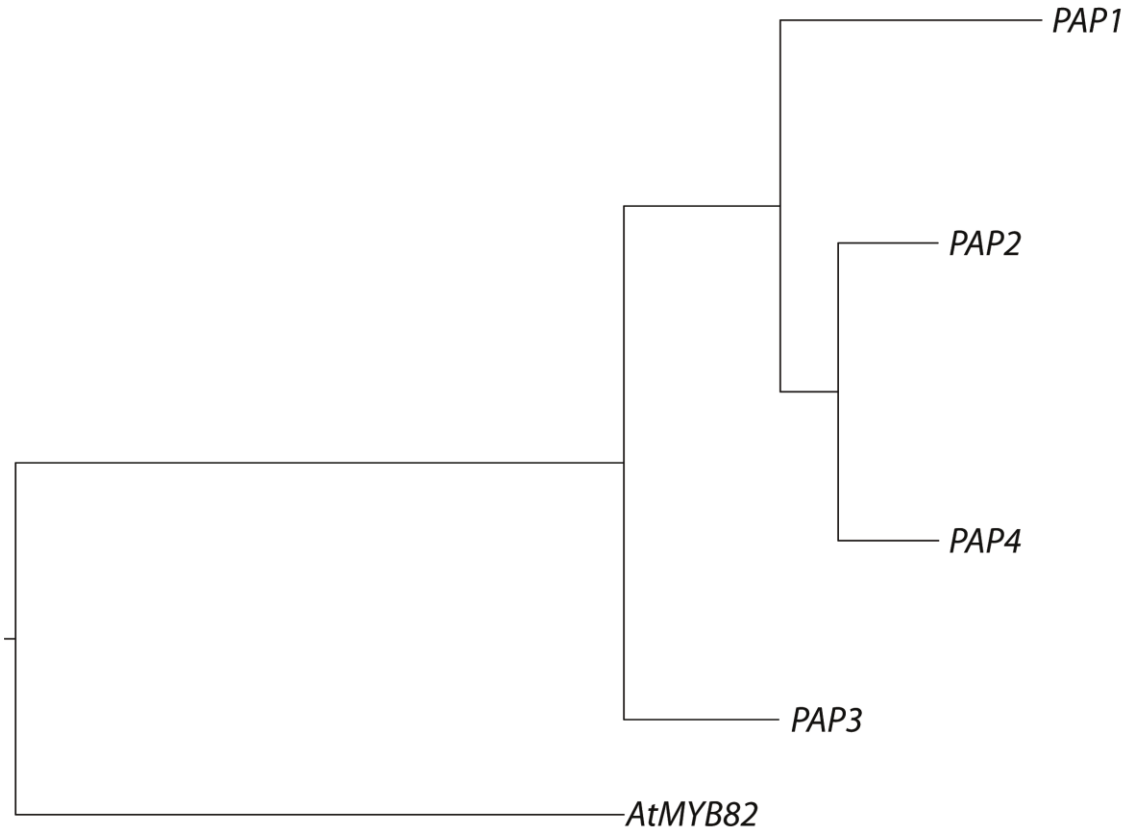


Figure 28 Bayesian phylogeny of consensus sequences of R2R3 MYB domain sequences of the *PAP* genes. As previously demonstrated in this work, the MYB regions of the *PAP* genes are highly conserved and are more likely to provide an accurate phylogeny by eliminating highly variable regions of the gene. *AtMYB82,* the most closely related *Arabidopsis thaliana* MYB gene to the *PAP* genes, was used as an outgroup. The gene sequence was taken from www.arabidopsis.org.

Given this result, we next determined the impact the 'undefined' regions of the genes have on inferred phylogenetic relationships. To do so, we removed the R2R3-MYB regions entirely from the genes and analysed only the remaining regions. The resulting tree produced a conflicting story to that of the R2R3-MYB region analysis, where *PAP2* and *PAP3* are shown to be the most recently diverged, *PAP2*, *PAP3* and *PAP4* are grouped together to the exclusion of all others and the *PAP* genes are grouped together to the exclusion of *AtMYB82* (*Figure 29*). In order to determine whether motifs within the genes could be used to heuristically identify and delimit genes within the R2R3-MYB gene family, we set about analysing previously proposed identifying motifs as well as attempting to identify novel motifs using our dataset.

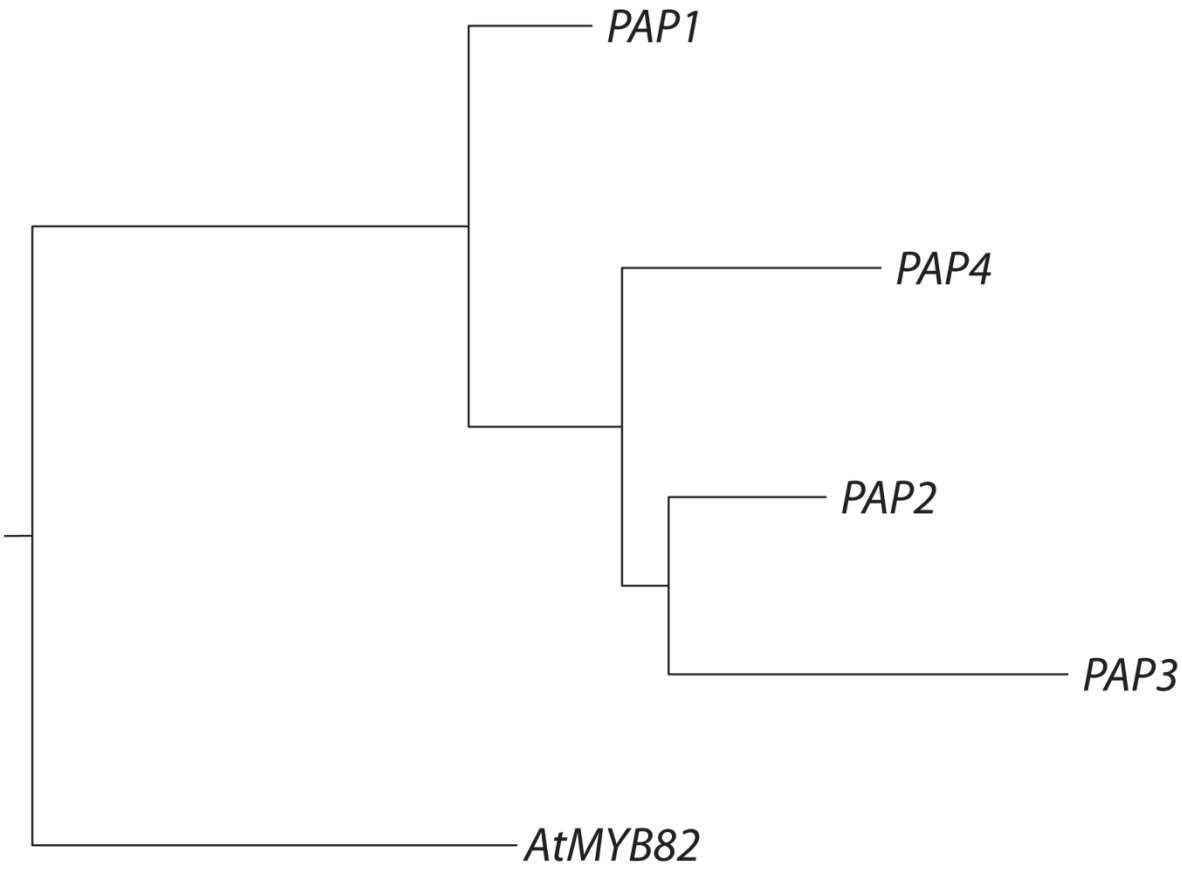


Figure 29 Bayesian phylogeny of consensus sequences of 'undefined' sequences of the *PAP* genes. The highly variable 'undefined' region of the *PAP* genes was analysed to determine whether it conflicts with the more conserved MYB domains in the *PAP* genes. *AtMYB82,* the most closely related *Arabidopsis thaliana* MYB gene to the *PAP* genes, was used as an outgroup. The gene sequence was taken from www.arabidopsis.org.

To further investigate the history of the *A. thaliana PAP* genes, we looked to the related species *A. lyrata* and *Brassica rapa*. In *A. lyrata,* we were able to identify five '*PAP*-like' genes, three of which were orthologous to *A. thaliana PAP* genes, *PAP1*, *PAP3* and *PAP2*. However, while there were two further *A. lyrata PAP*-like genes, neither were specifically orthologous to *PAP4.* In *B. rapa*, we identified a single locus orthologous to *PAP1, PAP2, PAP3* and *PAP4* (100% query coverage with 81% identity; 100% query coverage with 83% identity; 98% query coverage with 87% identity; and 100% query coverage with 82% identity to *B. rapa* locus HQ337792.1, respectively).

3.3.5 Linkage disequilibrium of the PAP genes

On its face, analysis of genetic linkage between the *PAP* genes confirmed what we expected, where *PAP1* is in linkage equilibrium (LE) to the other three physically linked *PAP* genes (*figure 30*). A

number of sites demonstrate a moderate value of R^2, though this is shown to be not significant and, in most cases, is the product of mutations being linked across the genes of a single accession. However, *PAP1* does demonstrate a high level of significant intragenic linkage disequilibrium (LD), related to the mutations defining the two predominant haplogroups. *PAP3*, *PAP4* and *PAP2* are located within 12 kb of each other, though they did not display a high level of LD, much less so than the intragenic LD of *PAP1.* There are a number of sites across the three physically linked *PAP* genes which appear to be in significant LD ($P < 0.01$), though the majority of these instances appear coincidental rather than legitimately linked. Where a number of accessions carry a particular mutation in one gene and have an apparently linked mutation in another, these mutations do not necessarily occur in combination, as a number of other accessions have either of the alleles independent of the other. We did observe an instance of significant LD ($R^2 = 1$; $P < 0.0001$) between two alleles of *PAP2* and *PAP4*. The nine accessions which carry the E209G amino acid replacement in *PAP2* also all carry the early stop codon form of *PAP4* (K140STOP). The E209 mutation in *PAP2* likely results in a coil rather than an α-helix in the tertiary structure of the protein. No accessions have the E209G mutation in *PAP2* without also having the K140STOP replacement in *PAP4*. However, the tenth accession with the K140STOP replacement, Sp-0, did not have the corresponding E209G mutation in *PAP2*, but rather maintained the majority *PAP2* allele with a glutamic residue at amino acid site 209.

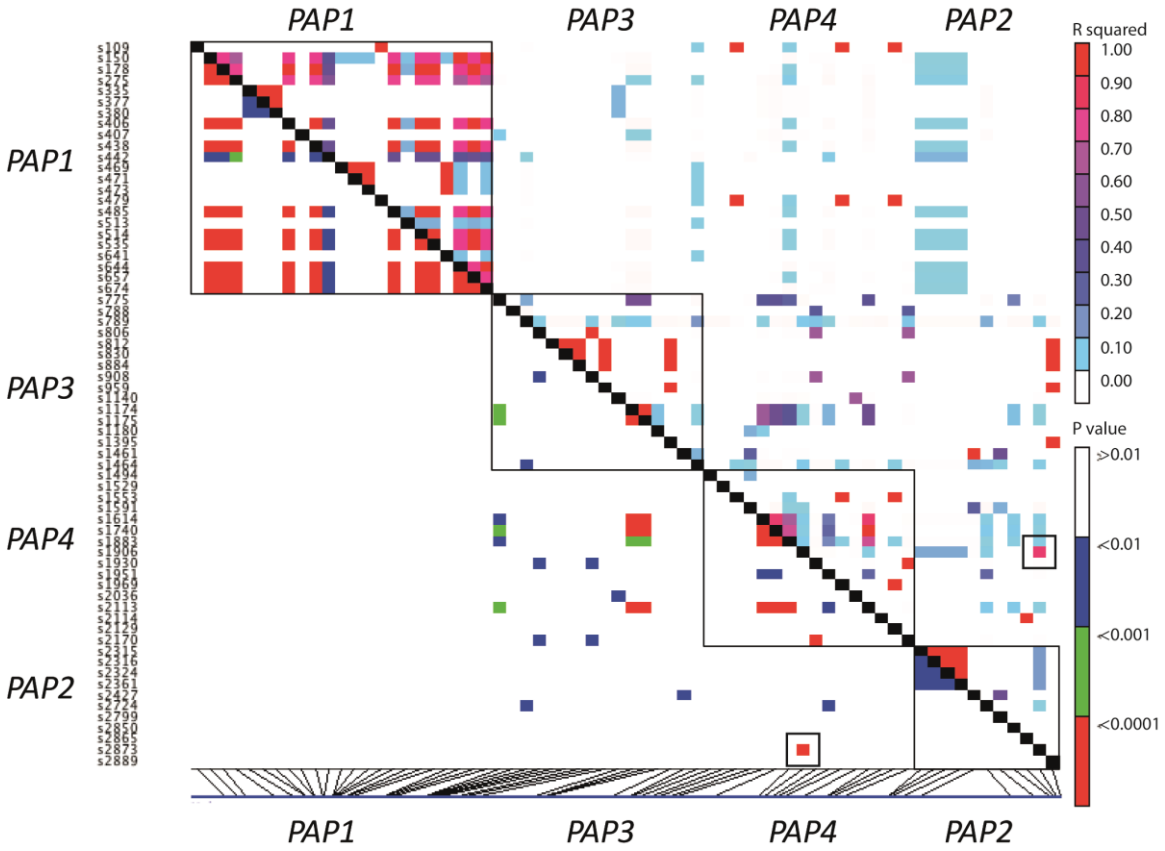


Figure 30 Linkage disequilibrium analysis of mutations of the *PAP* genes. Intragenic measures of linkage disequilibrium are shown boxed. The extent of linkage disequilibrium (R squared) above the black diagonal lne. The significance of any indication of linkage disequilibrium is tested and shown below the black diagonal line (P values). The nature and location in the concatenated sequences of the mutations is shown to the left of the figure. The mutations in demonstrating significant linkage disequilibrium (*PAP4*: K140STOP; *PAP2*: E209G) are shown in the small black boxes.

3.3.6 Unique motifs identifying the PAP genes

3.3.6.1 Motifs in the R2R3-MYB region. The demonstrably high level of sequence conservation makes the R2R3-MYB regions ideal for identification and delimitation of related genes. As such, the motif $[^{D}/_{E}]LX_2[^{R}/_{K}]X_3LX_6LX_3R$, located from amino acid sites 12 to 33 of the R3 repeat region, has been previously identified as useful for phylogenetic analysis of genes within the R2R3-MYB gene family, as it appears to be necessary for MYB-bHLH protein-protein interaction and is thereby relatively conserved within, but varied between, gene families (Zimmermann *et al.*, 2004), reflecting the integral role but varying targets of the R3 motif unique to each gene family. Across this 'R3 ID motif'

of 20 amino acids in the four *PAP* genes we analysed, only two sites differed from the PAP gene family consensus, one in *PAP1* and the other in *PAP3*, reflecting the recent divergence of the genes. To determine the efficiency and accuracy of delimitation using this sequence, we compared *PAP1, PAP2, PAP3* and *PAP4* to *AtMYB82.* As expected, the *PAP* genes were shown to share a common ancestor, with *PAP2* and *PAP4* were proposed to have the most recent common ancestor, and *PAP1* was proposed to be the most diverged of the *PAP* genes (*Figure 31*).

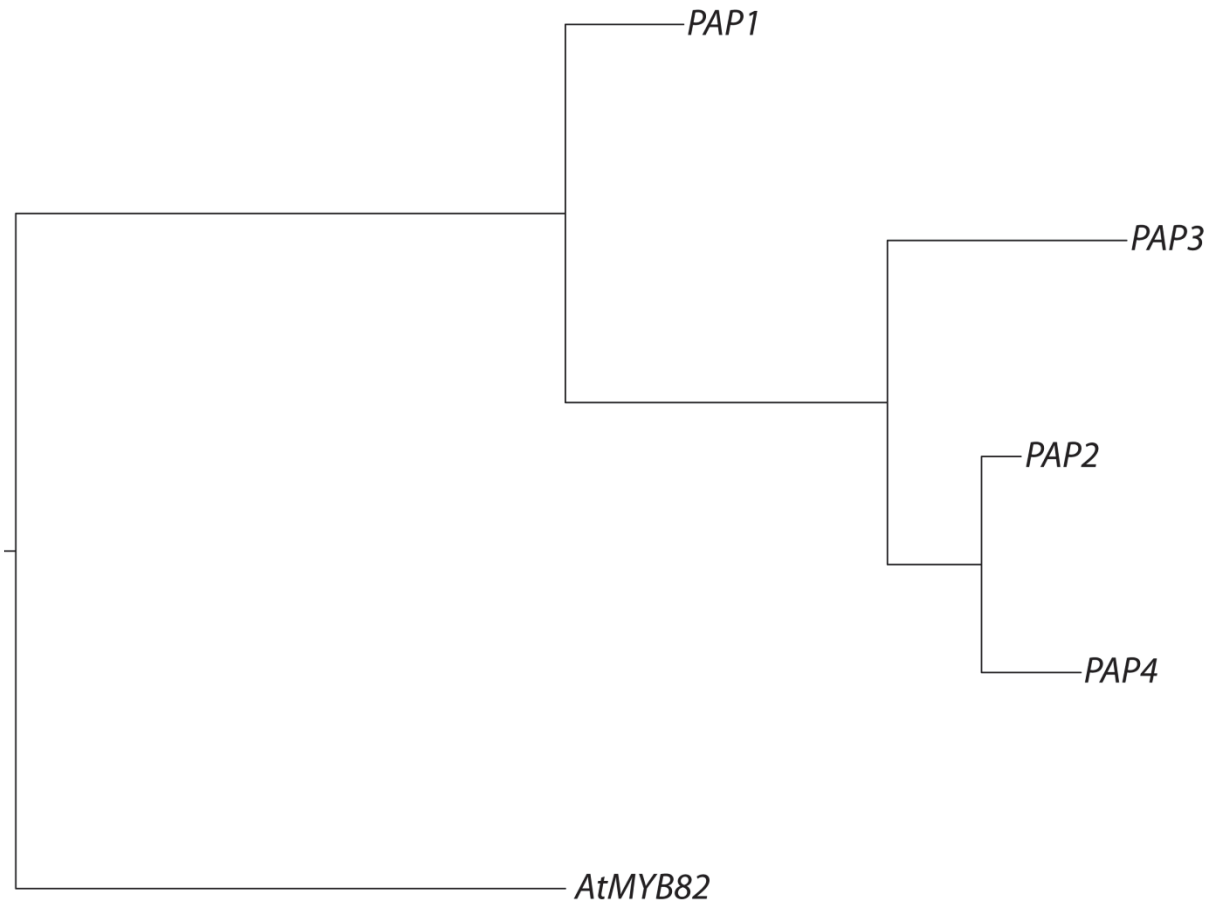


Figure 31 Bayesian phylogeny of consensus sequences of R3 'ID' motif of the *PAP* genes. The R3 'ID' motif is responsible for MYB-bHLH protein-protein interaction (Zimmerman *et al.,* 2004) and is therefore highly conserved, likely providing an accurate phylogeny of MYB genes. *AtMYB82,* the most closely related *Arabidopsis thaliana* MYB gene to the *PAP* genes, was used as an outgroup. The gene sequence was taken from www.arabidopsis.org.

To determine how much information is contained in this R3 ID motif relative to the remainder of the MYB domain, we performed a Bayesian analysis of the MYB regions of the PAP genes with the 20-residue long R3 ID motif removed, using *AtMYB82* as an outgroup. The resulting tree differed

from previous analyses again, with *PAP1* and *PAP4* shown sharing a common ancestor, *PAP1*, *PAP2* and *PAP4* grouped together to the exclusion of all others and the *PAP* genes grouped together to the exclusion of *AtMYB82* (*Figure 32*).

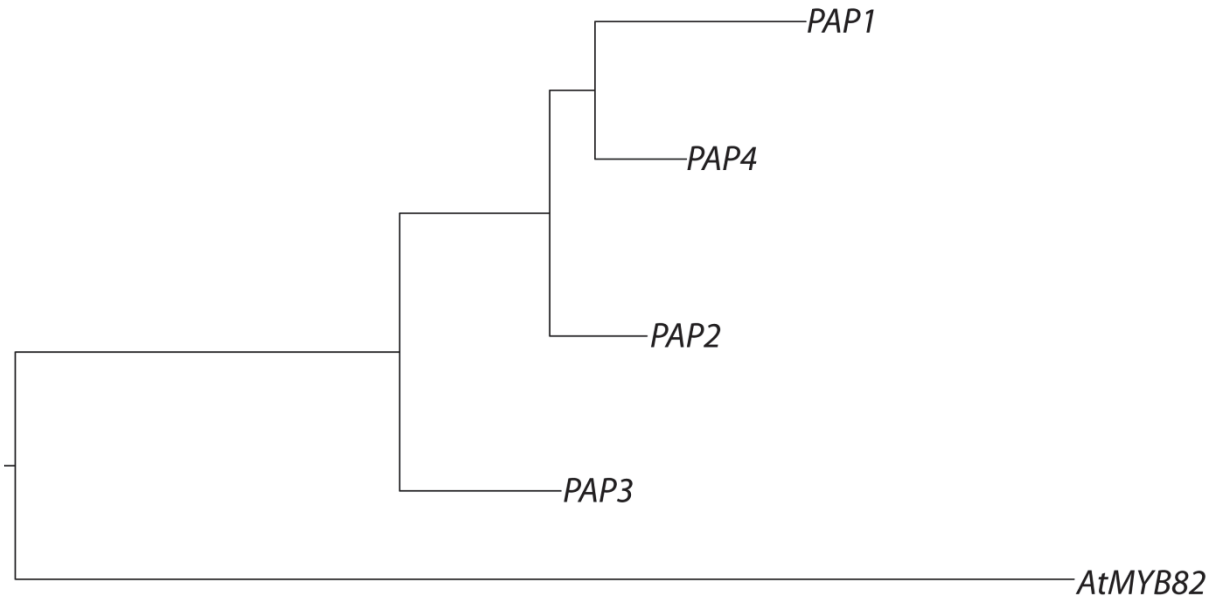


Figure 32 Bayesian phylogeny of consensus sequences of R2R3 MYB domain sequences of the *PAP* genes with the R3 'ID' motif removed to determine the level of unique information in the R3 'ID' motif compared with the remainder of the MYB domain. *AtMYB82,* the most closely related *Arabidopsis thaliana* MYB gene to the *PAP* genes, was used as an outgroup. The gene sequence was taken from www.arabidopsis.org.

3.3.6.2 Motifs in the undefined region. Kranz *et al.* (1998) undertook gene delimitation by motif identification in the *A. thaliana* R2R3-MYB gene family. They identified a motif in *PAP1* and *PAP2,* labelling the members 'subgroup 6'. The proposed motif shared by the two genes in the subgroup, VNNL[$^{M}/_{I}$][$^{N}/_{D}$]GDNWLE beginning at amino acid site 192, is maintained by *PAP1* and *PAP2* in all accessions we analysed, except for *PAP2* in accessions Ita-0 and Rld-2, where VNNLMNEDNMWLE and VNNLMNGDKMWLE were found, respectively (*the deviating residues are underlined*). *PAP3* and *PAP4* were not included in the Kranz *et al.* (1998) analysis as this study predated the discovery of these genes. Our analysis showed *PAP3* and *PAP4* do not display convincing evidence that this motif is being maintained.

Stracke *et al*. (2001) included PAP3 in their analysis of R2R3 MYB gene motifs. A second motif (KPRPR[$^{S}/_{T}$]F, beginning at site 140) was identified, linking *PAP1, PAP2* and *PAP3*. All accessions we analysed maintained this motif in *PAP3*, except for Alc-0 (KP---GSF), Sap-0, Bl-1, Ms-0 (KPR--SF), Bur-0 and Nok-1 (KPKPRSF). Stracke *et al*. (2001) also included *PAP4* in their analysis which, though phylogenetically closely related, was not found to share this motif. In this study, we found 38 of the 48 accessions analysed did maintain a similar motif at the corresponding site motif (KPRPRFF). However, ten of the accessions in our analysis, Enkheim-T, Gre-0, Jm-0, Mt-0, N7, Oy-0, Pa-1, Sav-0, Sp-0, Yo-0, had a non-synonymous mutation in the first codon of the motif, which alters the lysine residue to a stop codon. Col-0, the accession used by Stracke *et al*. (2001) in their study, also carries this mutation. Considering this, *PAP4* can be included in subgroup 6 of the R2R3 MYB gene family, based on the modified motif [$^{K}/_{*}$]PRPR[$^{S}/_{F}$]F. However, provided the truncated *PAP4* genes remains functional, we would expect the sequence upstream of the new stop codon to be maintained while the sequence downstream (including the currently conserved motif) to be released from selective constraint and likely deviate from consensus. In this case, this motif would not be a reliable identifying feature of the *PAP4* in the future.

3.3.6.3 De novo motif identification. We analysed 189 *A. thaliana* contiguous sequences to determine whether a common motif could be found linking the four PAP genes across all accessions. Predictably, all motifs initially identified were found in the R2R3-MYB regions; these regions were removed for further analysis to determine whether any conserved motifs could be identified outside of the R2R3-MYB regions to link the PAP four genes. The proposed motifs demonstrated low conservation; of the five motifs returned, three had sites that allowed up to four residues to produce a consensus sequence. The motif identified by Stracke *et al*. (2001) was included in part of the first motif returned for our data set, though the first site allowed up to three residues at the locus, resulting in the 'consensus sequence' [$^{K/}P_{/R}$][$^{P}/_{L}$][$^{R}/_{K}$]PR[$^{S}/_{F}$]F. The motif showing the greatest conservation across our analysed sequences in all four genes is located at the immediate 5' end of

the protein sequence: (M[$^{E}/_{G}$][$^{G}/_{E}$]S[$^{S}/_{P}$]KGH). We compared this motif to four *A. thaliana* R2R3 MYB genes, *Gl1*, *WER*, *AtMYB2*, *AtMYB82* and *AtMYB123*, none of which share this motif.

3.3.7 The PAP genes and transcriptional regulation

We sought to investigate whether the PAP genes experience similar regulatory context or are under the control of varying influences potentially resulting in different timing or function. Previously, the genes *WER, GL1* and *AtMYB23* have been shown to encode functionally equivalent genes which maintain similar transcriptional regulator motifs, differing in their timing and location of activation (Lee & Schiefelbein, 2001; Kang *et al*., 2009). The motif necessary for transcriptional activation of *WER* has been previously found to be located in the c-terminal domain of the gene, with similarly conserved regions in *GL1* and *AtMYB23* (Lee & Schiefelbein, 2001; Kang *et al*., 2009). In this study, we found that the *PAP* genes demonstrated intergenic conservation of the c-terminal domain also, hinting at the possibility that our candidate genes also maintain c-terminal domains for transcriptional activation. We noted that the c-terminal domain of *PAP3* differed considerably from the other three genes; when translated, *PAP3* differed from a consensus sequence of *PAP1, PAP2* and *PAP4* by 12 of 23 amino acids in this region.

Evidence of transcriptional down-regulation of genes in the R2R3 MYB family by siRNAs and miRNAs has been previously proposed for *PAP1*, *PAP2* and *PAP3* (Rajagopalan *et al.*, 2006). The locus *TAS4* gives rise to a 21bp siRNA which targets *PAP1*, *PAP2* and *PAP3* at sites directly overlapping the previously mentioned 'subgroup 6' motif KPRPRSF. This site is identical in *PAP1* and *PAP2*; eight base pairs into this region, *PAP3* differs by one nucleotide, where an adenine nucleotide is replaced with a guanine nucleotide. *PAP4* is also possibly targeted at this site, as it differs from *PAP3* by a single nucleotide only, where a cytosine nucleotide has been replaced by a thymine nucleotide at the 16th site of the motif sequence. miRNA828 is proposed to target *PAP3* at a 22bp segment of the gene towards the 3' end of the R3 repeat region (Rajagopalan *et al.*, 2006). The regions corresponding of

this target site differ by one base pair each in *PAP1* and *PAP4*, three base pairs each in *WER* and *PAP2*, and by four base pairs in *GL1*.

3.4 Discussion

3.4.1 Variation and selection between the PAP genes

Typically, in the *PAP* genes we observed exon 1 and the first half on intron 1 as having relatively low nucleotide diversity, increasing towards the end of intron 1 (*Figures 2-8*). We noted a departure from this trend in *PAP3*, where π was at its highest in exon 1 while maintaining low diversity through the entirety of intron 1. Tellingly, intron 1 of *PAP3* is greatly reduced in length compared to the other PAP genes. Assuming our predictions regarding relative ages of the PAP genes are correct, *PAP3* being the oldest in the lineage and *PAP4* being the youngest, it appears reduction in the length of intron 1 reflects the age of the gene; *PAP3* has by far the shortest intron 1 at 145 bp long, *PAP1* and *PAP2* are longer at 552 bp and 522 bp, respectively, while *PAP4*, what we predict to be the youngest of the four genes, has the longest at 613 bp. Given the reduced nucleotide diversity towards the beginning of intron 1 of *PAP1*, *PAP2* and *PAP4* compared to the end, we expected that the shorter intron 1 found in *PAP3* would align with the first half of intron 1 in *PAP1, PAP2* and *PAP4*. However, this was not the case; intron 1 of *PAP3* aligns with the second half of intron 1 in the other PAP genes.

We observed a variety of estimates of π across the coding regions of the five genes analysed. None of our genes demonstrated consistency with genome-wide estimates of π for *A. thaliana* (π = 0.0047; (Bakker *et al.*, 2006). The estimates of π for all our genes fell below the genome-wide estimate (*PAP2* and *PAP3* were less than half; *WER* was 10-fold lower), except *PAP1* (π = 0.0096). However, when the haplogroups of *PAP1* are measured separately, rather than measuring the entire *PAP1* dataset together, each was found to be well below this value (π (P1A) = 0.0015; π (P1B) = 0.0013). Generally, the *PAP* genes demonstrated low nucleotide diversity across the majority of exon 1 and part of exon 2, corresponding to the R2R3 MYB domain, interrupted by increased nucleotide diversity through intron 1. *PAP3* bucks this trend with the highest incidence of nucleotide diversity

located in exon 1. There was a surprisingly high level of nucleotide variation in intron 2 of *PAP4*; while the gene maintained the trend of the much shorter intron 2 displaying considerably less nucleotide variation than its respective intron 1, peaks of variation observed in intron 2 of *PAP4* (π = 0.025) exceeded the highest peaks observed at any site throughout the other *PAP* genes.

3.4.2 Variation and selection within the PAP genes

Selection pressure has not acted uniformly on the coding regions of the genes. Pair-wise *Ka/Ks* comparisons revealed a conserved pattern of selection in the *PAP* gene family, as well as in comparisons of the *PAP* genes with *WER*, where negative selection was evident through the R2R3 MYB domain and selective constraint was released in the 'undefined' domain where measures of *Ka/Ks* indicate neutral evolution is the most common state (*Figure 20-26*). This finding was not unexpected, highlighting the importance of the MYB domain for protein stability and protein-protein and protein-DNA interactions. Incidences of positive selection in the R2R3 MYB domain occur mostly either at the start of either the R2 or R3 regions of the MYB domain, as sequence changes towards the ends of these regions are more likely to affect protein folding and function (Dubos *et al*., 2010).

We observed areas throughout the coding sequence in the 'undefined' region which demonstrated low π intragenically, though intergenic conservation was not apparent. The last 70bp of the coding region demonstrated low levels of diversity both between and within genes, except for *PAP3* which has six intragenic SNPs in this area, one of high frequency. The c-terminal domain in *WER* demonstrates low π, inasmuch as *WER* demonstrates low levels of π across the entirety of the coding region, having a total of four SNPs underlying three peaks of nucleotide diversity. Conservation of this region in *GL1* and *WER* has previously been reported and has been shown to be necessary in a transcriptional regulatory role (Lee & Schiefelbein, 2001; Bloomer *et al*., 2012). This led us to suspect the presence of a C-terminal domain acting as a *cis*-regulatory element in the *PAP* genes. While *PAP1, PAP2* and *PAP4* share a similar c-terminal sequence, *PAP3* differs substantially. Even though *PAP1* is sufficient to regulate anthocyanin production and is predominantly expressed

under normal developmental conditions, *PAP2*, *PAP3* and *PAP4* demonstrate functional redundancy with *PAP1* and are able to compensate *pap1* mutants (Gonzalez *et al*., 2008). Nucleotide variation is low at the putative c-terminal domain compared to the rest of the 'undefined' domain, though *Ka/Ks* analyses did not suggest purifying selection was occurring in this region. Given the previously established differences in timing and location of the *PAP* genes (Lea *et al*., 2007; Muller *et al*., 2007; Gonzalez *et al*., 2008; Lillo *et al.*, 2008; Rowan *et al*., 2009; Shi & Xie, 2010), this is not unsurprising, as the genes would be expected to maintain intragenic consensus at the putative c-terminal domain while differing from each other in reflection of their differing expression contexts.

Teasing apart the timing and location of *PAP* gene expression has proven difficult due to the tight linkage of *PAP2* (At1g66390)*, PAP3* (At1g66370) and *PAP4* (At1g66380) on chromosome 1 (Gonzalez *et al*., 2008). Reduced gene expression of all four *PAP* genes driven by RNAi was shown to result in greatly reduced anthocyanin production. Our analysis of selection via pair-wise alignments indicated that *PAP3* has experienced selective constraint in imperative functional regions of the gene, suggesting *PAP3* is being conserved. Gonzalez *et al.* (2008) also showed that *PAP3* retains the capacity to recover anthocyanin production in *pap1* mutant lines using a constitutive expression construct, indicating functional integrity of the protein-protein and protein-DNA interaction regions of PAP3 is retained. However, the constitutive expression construct disregards the endogenous expression context. Further, truncated proteins with the c-terminal region missing have previously been shown to reduce or impair gene activity (Gonzalez *et al*., 2008; Velten *et al.*, 2010); for example, overexpression of the Col-0 form of *PAP4* producing a truncated protein with the c-terminal region missing results in inhibition of TTG1-dependant epidermal cell fate pathways, including anthocyanin production, while the full length L*er* PAP4 protein is capable of replicating the *PAP3* overexpression phenotype (Gonzalez *et al*., 2008). It is therefore possible that *PAP3* is regulated in a different context from *PAP1*, *PAP2* and *PAP4,* reflected in the diverging sequence in the putative c-terminal domain, while still being functionally equivalent to the other three PAP

genes. Ideally, reduction in expression of each *PAP* gene separately could be used to determine the timing and locality of expression of each gene.

3.4.3 Mutations affecting the MYB domains

Interestingly, while we observed polymorphisms in the R2 regions of all the *PAP* genes the only mutations observed in the R3 region of the MYB domain were found in *PAP3* (*Table 2.4*), both of which we suggest would result in protein loss-of-function due to the location and nature of the mutation (*Figure 15*). It is difficult to conclude that the *PAP3* locus in *A. thaliana* has experienced relaxed selection pressure, as only two accessions from our dataset carry mutations in *PAP3* at all; nor is it likely a result the expression context of *PAP3* having been rendered unnecessary due to the environment, as the accessions carrying the mutations, Sah-0 and Rld-2, are located in western Russia and Spain, respectively, and there is no evidence to suggest that other accessions in nearby locations have experienced relaxed selection at the *PAP3* locus. The poly-A insertion in a key coding region of Sah-0 has likely rendered this gene non-functional; if this is the case and selective pressure has been relaxed, it would be expected that upstream mutation would be occurring at a similar rate, as the MYB region would no longer have pressure exerted on it to maintain sequence integrity. However, this is not the case. While a number of mutations are present downstream of the insertion, including six non-synonymous SNPs unique to Sah-0, only two mutations are located upstream of the insertion, and one of these is shared by 16 other accessions which do not appear to have experienced relaxed selective pressures, at least in the MYB domain. The most likely scenario we propose is the G7W replacement unique to Sah-0 has resulted in relaxed selective pressures across the entirety of the *PAP3* gene resulting in the increased occurrence of unique mutations downstream, including the poly-A insertion.

3.4.4 Phylogenetic relationships between the PAP genes

Members of the PAP gene family have been shown previously to be redundant (Stracke *et al*., 2001; F. Zhang *et al.*, 2003; Gonzalez *et al*., 2008) and are clearly the result of recent duplications. In

an attempt to shed light on the evolutionary history of the PAP genes, we performed phylogenetic analysis using the demonstrably conserved MYB domains which suggested that *PAP2* and *PAP4* are the most recently diverged genes in the PAP gene family (*Figure 28*). However, the comparative age of *PAP1* and *PAP3* proved difficult to determine, as we next preformed phylogenetic analysis using the conserved 'R3 ID' bHLH interaction motif. This motif has been previously proposed for analysis of the relationship between the *A. thaliana* R2R3 MYB gene as it has been identified as the site of interaction between bHLH and MYB proteins and is therefore unique to each MYB type (Zimmermann *et al.*, 2004). Analysis using this motif suggested that *PAP1* was the oldest of the four genes, rather than *PAP3*, though the *PAP* genes were found together to the exclusion of *AtMYB82* (*Figure 31*) Using the R3 ID motif for identification and delimitation of R2R3-MYB genes appears to be a practical and sufficiently accurate method of rapid identification of functionally similar genes within the R2R3-MYB gene family. However, to better resolve relationships within gene families and between genes of a similar function, more data is required. Interestingly, we were able to identify five '*PAP*-like' genes in *A. lyrata,* one each orthlogous to *PAP1, PAP2* and *PAP3* though none orthologous to *PAP4;* the two other *PAP*-like genes were more closely related to *PAP2* and *PAP3* than all others, though orthologous to neither. Based on this evidence, it is attractive to conclude that the divergence of *PAP4* from *PAP2* occurred after the *A. thaliana-A. lyrata* split, estimated to have occurred ~5 million years ago (Charlesworth & Vekemans, 2005). However, it is possible *PAP4* diverged prior to the *A. thaliana-A. lyrata* split and was subsequently lost in *A. lyrata* or either of the non-orthologus *PAP*-like genes have since undergone sequence change to the point of no longer resembling the *A. thaliana PAP* genes.

Possibly narrowing the PAP gene duplication timeframe in the lineage to which *A. thaliana* belongs, we identified a single locus in *B. rapa* which is orthologous to the four *A. thaliana* PAP genes. The *Brassica-Arabidopsis* split has been previously estimated to have occurred 20-40 million years ago (Yang *et al.*, 1999; Koch *et al.*, 2001; Blanc *et al.*, 2003; Ziolkowski *et al.*, 2006). With this in mind, it is difficult to determine whether *PAP1, PAP3* and the common ancestor of *PAP2* and *PAP4*

arose previous to the *Brassica-Arabidopsis* split and were subsequently lost in the *Brassica* lineage but maintained in *Arabidopsis*, or whether they occurred after the *Brassica-Arabidopsis* split.

3.4.5 Allele association between the PAP genes

Given the *PAP1* locus is located more the 3 mbp from the other three *PAP* loci, a state of linkage equilibrium is expected and indeed found. Intragenic linkage disequilibrium (LD) was observed in *PAP1*, which is again expected as it reflects the seven mutations carried by nine accessions differentiating the two *PAP1* haplogroups (*Figure 30*). Still, between the three physically linked *PAP* loci, we did not observe as much LD as we expected. The three genes are located within 12 kb, which just exceeds the distance at which LD is expected to decay in *A. thaliana* (Kim *et al*., 2007). There are several alleles which appear to be in LD, though further investigation determined that these alleles are not linked, as a number of them were found to occur in accessions separate from each other, though coincidence would have it that some do occur in the same accessions. The sites we did observe demonstrating significant LD ($R^2 = 1$; $P < 0.0001$), E209G in *PAP2* and K140STOP in *PAP4*, allow for an interesting case study to determine whether the linkage between the two sites is affecting expression of function of these genes in some capacity, as we found nine accessions in our dataset maintaining both alleles, 35 maintaining the majority alleles (a glutamic acid residue at *PAP2* amino acid site 209 and a lysine residue at *PAP4* amino acid site 140), but only one accession, Sp-0, which has the early stop codon in *PAP4* without the apparently linked mutation E209G in *PAP2*. At the site corresponding to the E209G amino acid site in *PAP1* and *PAP3,* the majority allele finds a glutamic acid residue at this site, and in *PAP4* it is an aspartic acid residue, both of which fall into the same functional class and result in an α-helix in the protein tertiary structure rather than a coil produced resulting from the E209G replacement.

It is possible that these sites are linked as a result of the accessions having accumulated mutations in the refugia of *A. thaliana* during the pleistociene glaciation period and subsequently maintained these mutations once expansion of the separate populations occurred after glacial

retreat (Sharbel *et al*., 2000; Beck, Schmuths, & Schaal, 2008). However, if this were the case, we would expect to see more sites linked in these accessions at least, and likely a number of others. However, we did not observe this. It therefore seems likely that these two mutations are genetically linked rather than coincidently still being maintained after post-glacial expansion, though the significance or effect of such is difficult to determine based solely on protein structure prediction. Comparing expression and phenotypes of the accessions which maintain the majority forms of the *PAP2* and *PAP4* loci with the nine accessions which maintain the two minority alleles and Sp-0, which maintains the early stop codon in *PAP4* as well as the majority *PAP2* allele, would be an interesting study to determine how the two forms of each gene affect each other.

3.4.6 Identifying MYB genes using motifs

While it is attractive to use small sections of genes to quickly identify related genes, it is not always an accurate method for identifying how these genes are related. Here we used a number of previously suggested motifs located both in the MYB domain and the 'undefined' domain, as well as testing motifs we identified using our own dataset. For the most part, while these motifs could be used for grouping the *PAP* genes to the exclusion of other MYB genes, using them in attempt to establish divergence history proved to confound the matter. Simply, there does not appear to be enough information in a small section of a gene to accurately define how they are related. Where more data is available, such as in this case, a more accurate picture can be drawn. Still, while it is difficult to determine which sections of a gene should be used for identification and delimitation, logic dictates that the most conserved regions would be used, the same logic upon which motif identification is based. For this reason, we suggest the MYB domain is the most suitable region of the gene to provide an accurate answer to the question of evolutionary history. Based on this, we here propose that *PAP2* and *PAP4* are the most recently diverged of the *PAP* genes, while *PAP3* is the oldest (*Figure 28*).

3.4.7 Biallelic patterns of the PAP genes

The 48 accessions of *A. thaliana* analysed for PAP1 here reveal eight coding region alleles (haplotypes). The eight haplotypes fall into two predominant haplogroups based on seven common polymorphisms in the second and third exons, six of which are nonsynonymous (*Figure 10*). This manner of gene dimorphism has been reported for other members of the epidermal cell fate pathway, such as *GL1* (Hauser *et al.*, 2001; Bloomer *et al*., 2012) and several other *A. thaliana* loci (Innan *et al*., 1996; Kawabe *et al.*, 1997; Aguade, 2001; Tian *et al.*, 2002; Mauricio *et al*., 2003; Rose *et al*., 2004). Given the proposed history of restriction of *A. thaliana* populations to glacial refugia during the Plesitocene and recolonization post-glacial retraction (Sharbel *et al*., 2000; Beck *et al*., 2008), this observation is not surprising. *Ka/Ks* analysis of the two haplogroups of *PAP1* revealed purifying selection across the majority of the coding region with intermittent peaks of positive selection reflecting the polymorphisms underlying the haplogroup distinction. Differing environmental conditions of refugia during glaciation periods likely resulted in positive selection acting on the predominantly expressed *PAP1*. Interestingly, we also observed a similar biallelic pattern in *PAP4*, where 10 accessions were distinct from the other 38 accessions in our dataset based on 12 high frequency mutations, though only three accessions have both of the less common alleles. Only two of the mutations in *PAP4* confer changes to amino acid sequence (P132L, R205G), both of which are located in the 'undefined' domain of the coding sequence.

The discovery of biallelic variation in *PAP4* possibly narrows down the likely timeframe of the duplication origin of *PAP4*, placing it before the restriction of *A. thaliana* populations and subsequent expansion throughout Europe from glacial refugia (~120 000 years ago) (Beck *et al*., 2008). This finding of biallelic variation in both *PAP1* and *PAP4* adds the PAP gene family to the growing number of genes involved in epidermal cell fate determination displaying this characteristic (Hilscher *et al.*, 2009; Symonds *et al.*, 2011; Bloomer *et al*., 2012); this biallelic pattern in these genes, as well as the potential for sub-functionalization afforded by the functional redundancy of the

PAPs, could account for much of the quantitative variation seen in epidermal cells, as Bloomer *et al.* (Bloomer *et al*., 2012) pointed out.

Our dataset provided an interesting insight into the effect of duplication on genes. The variation in the four PAP genes contrasted greatly with *WER*. While *WER* is functionally equivalent with *GL1*, it still maintains a unique and specific role in *A. thaliana* and therefore maintains sequence integrity (Lee & Schiefelbein, 2001). Contrasted with that scenario is the functional redundancy of the PAP genes which are demonstrably capable of carrying moderate sequence variation. The PAPs also showed a different pattern of mutation to that of *GL1*, for example, where the majority of polymorphisms in the MYB domain were located in the R2 region, rather than the R3 region (Bloomer *et al*., 2012). Expression analysis of the PAPs would provide further insight into the results of this duplicated past, as it would determine whether the observed sequence variation has resulted in sub-functionilisation in the case of *PAP1* and *PAP3*, how the mutations we observed and deemed likely to result in non-functional copies of the genes *PAP3* and *PAP4* in some accessions affect anthocyanin accumulation and how the polymorphisms in the MYB domains have an effect on protein function.

4. An Investigation of the Genetic Architecture of Anthocyanin Accumulation

4.1 Introduction

Anthocyanins are the product of a branch of the flavonoid biosynthetic pathway, involving more than ten structural genes (Martin *et al.*, 1991; Springbob *et al.*, 2002). Regulation of anthocyanin accumulation and biosynthesis is the work of the *TTG1* regulatory network. This network involves the formation of a regulatory protein complex involving the eponymous WD-40 repeat motif TTG1, one of three bHLH genes, *GL3, EGL3* or *TT8*, and one of four R2R3 MYB genes, *PAP1*, *PAP2*, *PAP3* or *PAP4* (Zhang *et al.*, 2003). While expression of PAP2, *PAP3* and *PAP4* has been observed, it is generally accepted that *PAP1* is sufficient to complete the protein complex involved in regulation of anthocyanin accumulation and is the predominantly expressed gene of the four *PAP* genes (Gonzalez *et al.*, 2008; Shi & Xie, 2010). Anthocyanin accumulation is considered a generic stress response in plants, occurring under a range of conditions including during the senescent stage of the plant life cycle (Powles, 1984; Hoch *et al.*, 2001). As there are a number of factors involved in anthocyanin biosynthesis, and anthocyanin accumulation is instigated under a range of conditions, there is a wide array of variation in the timing and extent of anthocyanin accumulation observed in natural populations of *A. thaliana.*

Of the genes we know to be involved in anthocyanin biosynthesis and regulation of accumulation, the *PAP* genes are a likely source of genetic variation for variation in anthocyanin accumulation, as the other genes involved in the regulation of anthocyanin biosynthesis and accumulation are pleiotropic. The structural genes of the anthocyanin biosynthetic pathway are also involved in producing a range of flavonoids; *TTG1* and the bHLH genes of the *TTG1* regulatory complex are involved in the regulation of all epidermal cell traits in *A. thaliana*. Because of the range of traits these genes are involved in, mutations resulting in changes to protein structure and function would likely not be tolerated by the suite of traits they contribute to, and as such, would not remain in the

gene pool. The *PAP* genes, however, provide the specificity the TTG1 protein complex requires to regulate anthocyanin accumulation and are not involved in regulation of other epidermal traits. Further, as we have previously established in this work, there are a number of mutations already present in *PAP* genes found in natural accessions of *A. thaliana,* suggesting potential changes to protein structure and function. For this reason, we expect variation in anthocyanin accumulation will be associated with the *PAP* gene loci.

Traditional methods of analysing the genetic architecture of phenotypic variation rely on recombinant inbred line (RIL) populations derived from two parent accessions to determine the genetic involvement of a certain phenotype. Here, we seek to identify the loci underlying the variation observed in *A. thaliana* using a multi-parent advanced generation inter-cross (MAGIC) population (Kover *et al*., 2009), which is bred from a total of 19 *A. thaliana* accessions. The 19 accessions were chosen in an attempt to incorporate as much genetic variation from *A. thaliana* as possible. Mapping the genetic bases of anthocyanin content variation in this population allows us to (i) characterise the genetic architecture of the trait by determining how many loci are involved and what their relative effects are and (ii) to test the hypothesis that the *PAP* loci are involved. Our results identify several loci throughout the *A. thaliana* genome that are associated with variation in anthocyanin accumulation.

4.2 Materials and Methods

4.2.1 Plant material and growth conditions. We used 19 natural *Arabidopsis thaliana* accessions and the first panel of 527 recombinant inbred lines from a multi-parent advanced generation inter-cross (MAGIC) population (Kover *et al*., 2009). Four replicates per line were potted in a fully randomised layout in 4-cm cells of seed raising mix (Odering Nurseries, NZ) in 72-cell flats (Hummert International, Inc.), with approximately five seeds planted per cell. The flats were sprayed with 1.8g/L (w/v) Terrachlor fungicide solution and stratified at 4°C for 12 days. Flats were then moved to 20°C under 15-hr light and thinned to two to three plants per cell 7 days post-germination; a second round of thinning, leaving a single plant per cell, was undertaken 14 days post-germination. The flats

were placed in a plant growth room at ~25°C with a 15:9 h day:night cycle. 42 days after vernalization, the temperature was reduced to ~15°C and the day:night cycle was adjusted to 8:16 h to slow vegetative growth and encourage anthocyanin production. The growing conditions were reverted after 7 days under these conditions; after another 7 days, pigment extractions were undertaken as described below.

4.2.2 Pigment extraction and analysis. We initially used a pigment extraction protocol used in a previous study (Albert *et al*., 2009), though this method proved impractical for large scale extractions such as ours. Previous attempts using this protocol in a smaller RIL population Hi-0 x Ob-1 returned heritability values less than 0.5. We used a modified version of the methanol pigment extraction protocol used by Albert *et al.* (2009) to extract pigments from the two leaves considered the most red visually on each plant. Two leaf discs were taken from the tip and base of the leaf over the midrib using a standard hand-held hole-punch. The discs were placed 1ml 96-well plates (Thermo Scientific) and were sealed using silicon Axymat covers (Axygen). Pigment extraction used 500µL acidified methanol (0.1M HCl) and proceeded for 24 h at 4°C on an orbital shaker set to 1.5krpm in 96-well plates (Thermo Scientific) sealed with silicon Axymat covers (Axygen). Extraction occurred in the dark to minimise pigment degradation. After overnight extraction, 250 µL of each sample was analysed for both anthocyanins (ʎ = 530nm) and chlorophyll (ʎ = 657nm) using a PowerWave XS spectrophotometer (BioTek) in 96-well spectrophotometer plates. The anthocyanin content of each plant was calculated by offsetting chlorophyll interference on absorbance using the formula $A_{530} - (0.25A_{657})$ (Mancinelli *et al.*, 1988).

Statistical analysis. Single factor ANOVA of raw data reads from spectrophotometric analysis was performed using the Analysis ToolPak in Microsoft Excel 2010. Broad-sense heritability was calculated as V_g/V_p from ANOVA results. As there are few options for mapping QTL in multi-parent populations, the developers of the MAGIC population have generated a mapping pipeline (Wellcome Trust Centre for Human, Oxford, UK), which is available to users. The pipeline tests each of >10,000 markers for associations with trait variation and generates LOD plots across the *A. thaliana* genome.

Based on models run from empirical data, marks with LOD scores greater than four are considered statistically significant.

4.3 Results

4.3.1 Heritability and mapping of anthocyanin accumulation

The broad-sense heritability estimate for anthocyanin production in the MAGIC population of *A. thaliana* is 0.768. This is based on 406 lines (including 17 of the 19 parent accessions) from the MAGIC panel out of a possible 527, as a number of lines either did not germinate or were inviable upon collection. The results of the screen show a roughly 20-fold distribution for anthocyanin production with a normal to bimodal distribution (*Figure 33*). We observed significant QTL (logP > 4) on all *A. thaliana* chromosomes except chromosome 3 (*Figure 3.2, Table 8*). Of particular interest were peaks near obvious candidate genes. The *PAP1* locus was located in a peak of associated SNPs with the third highest logP scores (logP = 5.86) on chromosome 1 from 19750-23000 kbp (*PAP1* locus = 21233-21235 kbp). *PAP3, PAP4* and *PAP2,* which are physically linked back-to-back-to-back, were also found in a narrower peak of associated SNPs on chromosome 1, from 24750-25150 kbp, with a lower logP score (logP = 4.84) than that which the *PAP1* locus was found in. We also sought to identify whether there was an association between phenotypes and other genes known to be involved in regulation of anthocyanin accumulation, namely *GL3*, *EGL3*, *TT8* and *TTG1*, as well as the structural genes involved in anthocyanin biosynthesis (*Table 3.2*). There is no evidence for a link between phenotypic variation in anthocyanin accumulation and these regulatory and structural genes, as we found no associated SNPs at these loci.

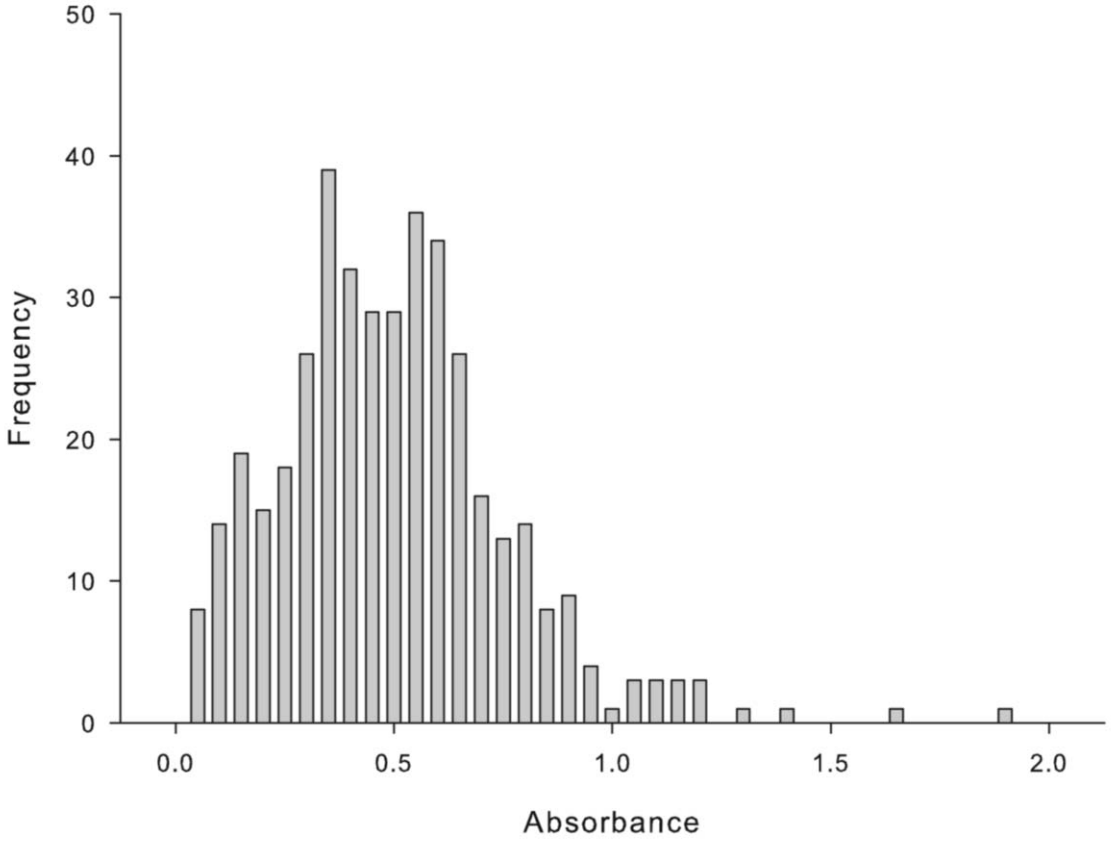


Figure 33 Distribution of measures of anthocyanin absorbance in the MAGIC population of *Arabidopsis thaliana.* The population demonstrates a 20-fold normal to bimodal distribution for anthocyanin production. Measures of absorbance are grouped in bins increasing in increments of 0.05 and plotted against frequency. n=406.

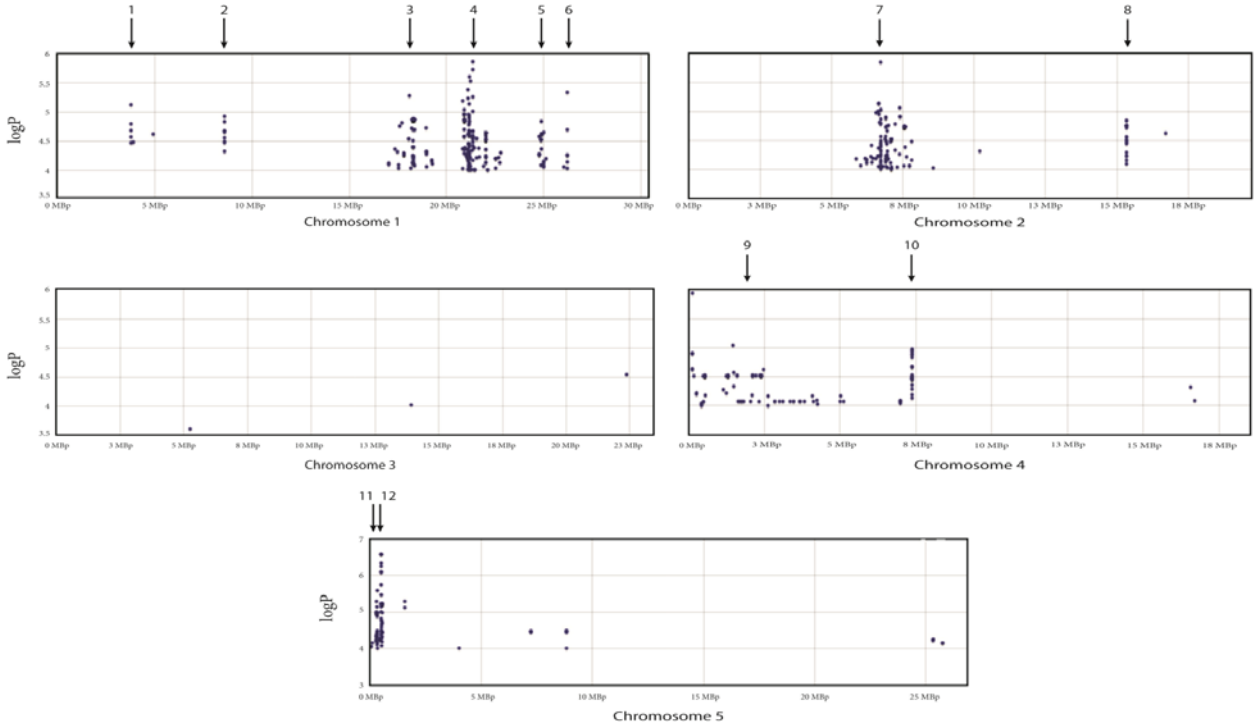


Figure 34 Chromosome maps of *Arabidopsis thaliana* with associated loci plotted against logP scores. Chromosome number is indicated below each plot. Each plot point denotes a positive association for a particular SNP marker with anthocyanin accumulation variation. Based on models run from empirical data, marks with logP scores greater than four are considered statistically significant The numbered arrows above the plots indicate the location of each peak (Table 8).

Table 8 Summary of the location and function of genes likely underlying loci associated with anthocyanin accumulation. The '/' between genes indicates that the associated loci was located between these two genes.

Peak	Chromosome	Location	Associated genes	Identity	LOD score
1	1	3763-3764 kbp	At1g11230	Uncharacterised protein	5.12
2	1	8576.25-8576.75 kbp	At1g24212	Psuedogene	4.92
3	1	17000-19500 kbp	At1g47128	Senescence-associated gene	5.28
4	1	19750-23000 kbp	*PAP1,* At1g58360	Regulation of anthocyanin accumulation, SAG	5.86
5	1	24750-25150 kbp	*PAP3, PAP4, PAP2*	Regulation of anthocyanin accumulation	4.84
6	1	26215-26217 kbp	At1g69690	Transcription factor TCP15	5.34
7	2	5750-8000 kbp	At2g15410/At2g15420	transposable element gene/uncharacterised protein	5.85
8	2	15333-15335 kbp	At2g36560/At2g36570	Uncharacterised proteins	4.84
9	4	0-3000 kbp	At4G01610	SAG	5.95
10	4	7370-7375 kbp	*TTPF/APT4*	Uncharacterised proteins	4.98
11	5	270-330 kbp	?	?	5.59
12	5	470-550 kbp	?	?	6.58

Table 9 Location and function of genes involved in regulation of anthocyanin accumulation and biosynthesis

Chromosome	Location	Gene Name	Identity
1	23599.65-23602.90 kbp	*EGL3*	Regulation of anthocyanin accumulation
3	19025.41-19026.58 kbp	*TT6* (F3H)	Anthocyanin biosynthesis
3	20430.11-20431.47 kbp	*TT5* (CHI)	Anthocyanin biosynthesis
4	6182.02-6186.49 kbp	*TT8*	Regulation of anthocyanin accumulation
4	12004.76-12006.20 kbp	*TT18* (LDOX)	Anthocyanin biosynthesis
5	2560.39-2563.10 kbp	*TT7* (F3'H)	Anthocyanin biosynthesis
5	4488.76-4490.03 kbp	*TT4* (CHS)	Anthocyanin biosynthesis
5	8370.70-8372.84 kbp	*TTG1*	Regulation of anthocyanin accumulation
5	16529.45-16532.87 kbp	*GL3*	Regulation of anthocyanin accumulation
5	17164.14-17165.91 kbp	*TT3* (DFR)	Anthocyanin biosynthesis

In total, we identified 12 QTL across the *A. thaliana* genome associated with our phenotype, including the two peaks associated with the *PAP* genes. We observed two regions of less than 5 kbp, one on chromosome 2 (15333-15335 kbp) and one on chromosome 4 (7370-7375 kbp), both with more than 20 SNP markers associated with our phenotype. The genes closest to these sites (AT2G36560/AT2G36570 and TTPF, CPuORF26/APT4, respectively) do not currently have a function ascribed to them. We noted another site of associated SNPs on chromosome 1 in a region of less than 500 bp, with the closest locus being a pseudogene. We next inspected a suite of senescence-associated genes (SAG) (Gepstein *et al*., 2003) to determine whether any of these were underlying the peaks of associated SNPs we observed in our analysis. We found two senescence-associated genes in regions with a number of associated SNPs. Also, we found two SNPs in close physical proximity to three other senescence related genes (*Table 10*). Despite pursuing a series of genes, there were no obvious candidates for QTL on chromosome 5.

Table 10 Location and function of senescence-associated genes

Chromosome	Location	Gene	Associated SNPs
1	4927.01-4928.69 kbp	At1g14400	1
1	10770.81-10775.4 kbp	At1g30460	0
1	12867.13-12868.9 kbp	At1g35160	0
1	17282.83-17285.67 kbp	At1g47128	peak three (*see table 8*)
1	20064.71-20068.47 kbp	At1g53750	0
1	21676.53-21680.5 kbp	At1g58360	peak four (*see table 8*)
2	8979.58-8981.23 kbp	At2g20860	0
2	9353.17-9354.65 kbp	At2g21950	0
2	17776.24-17777.72 kbp	At2g42690	0
3	925.61-930.97 kbp	At3g03720	0
3	3860.29-3863.05 kbp	At3g12120	0
3	5273.9-5275.1 kbp	At3g15580	1*
3	5330.32-5333.75 kbp	At3g15730	1*
3	6089.61-6091.23 kbp	At3g17790	0
3	16907.45-16909.99 kbp	At3g46000	0
3	19960.87-19963.95 kbp	At3g53920	0
3	20549.59-20552.21 kbp	At3g55430	0
4	694.69-697.20 kbp	At4G01610	peak nine (*see table 8*)
4	9171.46-9173.10 kbp	At4g16190	0
4	10103.99-10104.73 kbp	At4g18280	0
4	13861.71-13864.5 kbp	At4g27830	0
4	14819.19-14820.59 kbp	At4g30270	0
4	14881.85-14883.48 kbp	At4g30440	0
4	15900.27-15903.26 kbp	At4G32940	0
5	2709.84-2710.58 kbp	At5g08410	0
5	3685.08-3687.76 kbp	At5g11520	0
5	8500.47-8502.22 kbp	At5g24770	0
5	8507.59-8508.95	At5g24780	0
5	13108.15-13111.63 kbp	At5g34850	0
5	21643.89-21647.70 kbp	At5g53350	0
5	23346.76-23350.03 kbp	At5g57655	0
5	24279.89-24282.39 kbp	At5G60360	0

4.4 Discussion

Here, we were able to link *PAP* genes to the variation in anthocyanin accumulation using the MAGIC mapping population. There was a much stronger association of the phenotype with *PAP1* compared to the other three *PAP* genes, though this is unsurprising as there is a previously established predominance in expression of *PAP1*, though not to the exclusion of the other *PAP* genes (*Table 8*) (Gonzalez *et al.*, 2008; Shi & Xie, 2010). Conversely, we observed no association between our phenotype and the other genes involved in the anthocyanin accumulation regulatory complex, namely the bHLH genes *GL3*, *EGL3* or *TT8* and the WD-40 repeat motif *TTG1*. Again, this is not unexpected, as *TTG1* acts as the scaffold for a protein complex regulating all epidermal cell traits in *A. thaliana* (Zhang *et al.*, 2003). As such, we observe no natural knock-out mutations of *TTG1*. Similarly, the three bHLH genes are too involved in regulation of all epidermal cell traits, and thereby we do not see mutations which are altering the anthocyanin phenotype.

We examined a suite of senescence-associated genes as likely candidate genes underlying the peaks of associated SNPs we observed across the *A. thaliana* genome. We found two SAGs within these observed peaks, one on chromosome 1 and the other on chromosome 4. We found another SAG on chromosome 1 associated with a single SNP (logP > 4) and two more SAGs on chromosome 3 associated with a single SNP (logP > 4). Considering the age at which we harvested our samples, it is not unexpected that we should find senescence associated with our phenotype. Even after examining a number of candidate genes, we were still unable to identify good candidate genes underlying the peaks of associated SNP markers on chromosome 5.

While we were able to associate the candidate *PAP* genes with the anthocyanin accumulation phenotype, and another two peaks are potentially associated with senescence-associated genes, this only accounts for four of the 12 observed QTL. We were unable to definitively identify genes associated with anthocyanin accumulation underlying the remaining peaks. Given the wide range of circumstances under which anthocyanin accumulation is induced, there is a whole suite of influences

that potentially influence this phenotype. For this reason, we propose further experimentation incorporating an analysis of anthocyanin accumulation through a developmental series to determine whether the associated loci are predominantly senescence related, or some other genes are involved which we have not yet elucidated as being involved in anthocyanin biosynthesis and accumulation. Further, while it is satisfying to identify an association between the *PAP* genes and variation for anthocyanin accumulation, particularly the apparent predominance of involvement from *PAP1*, it would be an interesting exercise to subject a test population to varying conditions to determine whether *PAP1* is indeed capable and sufficient to regulate anthocyanin accumulation across all manner of conditions, or whether the duplicate *PAP* genes respond to different conditions.

A number of sites remain to be pursued for the purpose of identifying their nature and role in association with variation of anthocyanin accumulation. Still, our work supports the expectation that the *PAP* genes contribute to this observed variation, though this is not yet definitive. Further, we have potentially excluded both the structural genes of the biosynthetic pathway and the other genes of the *TTG1* regulatory complex from association with our trait, at least for the accessions included in the MAGIC mapping resource. This provides a stable point from which to venture into identifying other factors contributing to the natural variation of such a complex trait as anthocyanin accumulation, as well as insight into the nature and mechanism of natural variation in general.

5. Conclusion

In our consideration of the four duplicate *PAP* genes, we have addressed a number of issues regarding natural variation pertaining to anthocyanin accumulation. When considering the *PAP* genes in comparison to the *WER* locus, it seems a logical conclusion that duplication of the genes regulating anthocyanin accumulation has allowed for tolerance of more mutation than that which would be allowed for a highly pleiotropic gene, such as *TTG1,* or a gene which carries out single role alone, such as *WER.* Our consideration of the genetic architecture of the *PAP* genes in a number of accessions from the world-wide population of *A. thaliana* revealed an interesting history, where we observed a biallelic pattern across the *PAP1* gene, likely a result of population isolation in glacial refugia during the Pleistocene. The maintenance of these two predominant haplogroups presents an interesting case study to determine whether expression and function varies between the two groups as a result of these coding region differences. Conveniently, we found that *PAP1* in the accession Sakata has likely undergone recombination at some point in the past as it shares some identifying mutations from each haplogroup and would provide an intermediate between them when considering the effects these characteristic mutations have on their respective phenotypes.

While it has proven difficult to conclusively determine the ages of the *PAP* genes in relation to each other, we suggest it is most likely that *PAP2* and *PAP4* are the most recently diverged. Considering the extent of nucleotide diversity across the four *PAP* genes, it appears that the divergence event between *PAP1* and the other *PAP* genes resulted in *PAP1* being the primary MYB gene involved in regulation of anthocyanin accumulation in *A. thaliana*, at least in optimal growing conditions, as the two predominant haplogroups of *PAP1* revealed the lowest nucleotide diversity of the *PAP* genes. This has demonstrably been to the detriment of sequence conservation of the *PAP3* gene, as it revealed the second highest level of nucleotide diversity across the coding region of the four *PAP* genes. Further, *PAP3* is the only gene of the four duplicate genes where we observed non-synonymous mutations in the R3 region of the MYB domain. Comparing the divergence event

between *PAP2* and *PAP4* is not as simple. The gene duplication events likely occurred in quick succession and though it is intuitive to expect the most recently diverged genes to have the least nucleotide diversity of a group of duplicated genes, this is not the case. We find *PAP2* to have nucleotide diversity the second lowest while *PAP4* has the highest. Our suggestion that a c-terminal activation domain is present in the *PAP* genes more than likely excludes *PAP4* from being involved in regulation of anthocyanin accumulation in ten accessions from our dataset, as the early stop codon found in these accessions occurs upstream of the c-terminal domain. However, the impact of the observed link between the stop codon in *PAP4* and a non-synonymous mutation in *PAP2* merits investigation. Again, we conveniently found an accession, Sp-0, which would act as an intermediate study between the two genotypes, as Sp-0 maintains the early stop codon in *PAP4* without the associated mutation in *PAP2.*

We were able to demonstrate the preponderance of *PAP1* in its effects on natural variation of anthocyanin accumulation by mapping QTL in a multi-parent population, which we expected to find as it has been previously shown to be the predominantly expressed *PAP* gene (Zhang *et al.*, 2003). Variation in anthocyanin accumulation was not exclusive to the *PAP1* locus; we also observed a number of SNPs associated with the *PAP2, PAP3* and *PAP4* loci, though to a lesser extent. Finding a link between the anthocyanin accumulation phenotype and the *PAP3:PAP4:PAP2* locus presents compelling reason to determine whether the sequence conservation of *PAP2* we observed in our molecular analyses is due to the continued, though minor, involvement of *PAP2* in regulation of anthocyanin biosynthesis, and whether the extent of this involvement varies under different conditions. More so than the *PAP3:PAP4:PAP2* locus, we observed ten other loci spread across the *A. thaliana* genome associated with variation in anthocyanin accumulation. A number of these we deemed likely to be senescence-related while others did not have an obvious involvement in regulation of accumulation or biosynthesis of anthocyanins or even a definitive identity yet attributed to them. Two sites on chromosome 5 which had a number of associated SNPs indicating the locus is involved in the natural variation of anthocyanin accumulation covered a wide region of

the chromosome so that we were unable to conclusively pinpoint a candidate gene amongst the large number of genes present.

The contribution of one or more of the *PAP* genes other than *PAP1* is encouraging for further investigation, as their minor contribution to variation in anthocyanin accumulation under moderate environmental conditions in the least suggests they are functional and potentially have the capacity to regulate anthocyanin biosynthesis. Our findings also suggest that neither the remainder of the proteins involved in completing the TTG1 regulatory complex nor the structural genes of the anthocyanin biosynthesis pathway affect the observed variation in anthocyanin accumulation in the MAGIC population resource used. The exoneration of these loci from involvement in variation of anthocyanin accumulation narrows the focus of future work primarily on the *PAP* loci as the source of observed variation within the TTG1 regulatory network as well as a number of novel sites across the *A. thaliana* genome. Mapping studies using varying environmental conditions could elucidate whether increased stress upon the individual or particular conditions will result in differing weights of responsibility placed upon the four *PAP* genes for regulation of anthocyanin accumulation. Further, studies varying the timing of sample collection from the multi-parent population would likely determine the extent of involvement of the senescence-related loci and whether this is a constant involvement or indeed, as we would expect, is limited to the terminus of the *A. thaliana* lifecycle. A combination of these studies would likely also assist in narrowing the number of genes potentially underlying the peaks for which we could not identify a candidate gene. Provided further investigation does suggest that there is a consistent contribution from these loci to the variation in anthocyanin accumulation under a range of conditions, the identity of these associated genes and loci would be interesting as some have an even stronger association with our anthocyanin phenotype than *PAP1*.

Our work regarding the effects of gene duplication on sequence conservation and function of the *PAP* genes and the underlying causes of the natural variation observed in epidermal accumulation of

anthocyanins raised more questions than it addressed. Still, we have been able to narrow the scope within which we find the major contributors to variation in this trait. Our molecular analyses reveal considerable natural variation at the *PAP* loci, the majority of which is likely due to relaxed sequence conservation following gene duplication, while our QTL mapping analyses link these loci to the anthocyanin accumulation trait. Expression analyses have the potential to elucidate the effects this observed mutation has on anthocyanin accumulation, as well as the specific roles of the four duplicated genes. Much of the variation in epidermal traits across *A. thaliana* is likely due to allelic variation in the TTG1 regulatory network (Hilscher *et al.*, 2009; Symonds *et al.*, 2011; Bloomer *et al*., 2012). For anthocyanin accumulation, variation driven by the TTG1 regulatory network appears to be more limited to the *PAP* genes themselves, providing an interesting insight into the molecular basis for epidermal cell trait variation and paving the way for important work in the future.

6. References Cited

Aguade, M. (2001). Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in Arabidopsis thaliana. *Molecular Biology and Evolution*, *18*, 1–9.

Albert, N., Arathoon, S., Collette, V., Schwinn, K., Jameson, P., *et al.* (2010). Activation of anthocyanin synthesis in *Cymbidium* orchids: variability between known regulators. *Plant Cell Tissue and Organ Culture*, *100*(3), 355–360.

Albert, N., Lewis, D., Zhang, H., Irving, L., Jameson, P., & Davies, K. (2009). Light-induced vegetative anthocyanin pigmentation in Petunia. *Journal of Experimental Botany*, *60*(7), 2191–2202.

Alonso-Blanco, C., El-Assal, S., Coupland, G., & Koorneef, M. (1998). Analysis of Natural Allelic Variation at Flowering Time Loci in the Landsberg *erecta* and Cape Verde Islands Ecotypes of *Arabidopsis thaliana*. *Genetics*, *149*(2), 749–764.

Atkinson, D. (1973). Some general effects of phosphorus deficiency on growth and development. *New Phytologist*, *72*(1), 101–111.

Bakker, E., Toomajian, C., Kreitman, M., & Bergelson, J. (2006). A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell*, *18*, 1803–1818.

Bandelt, H., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biological Evolution*, *16*, 37–48.

Beck, J., Schmuths, H., & Schaal, B. A. (2008). Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology*, *16*, 37–48.

Beggs, C., & Wellmann, E. (2008). Analysis of light-controlled anthocyanin formation in coleoptiles of *Zea mays* L.: The role of UV-B, blue, red and far-red light. *Phytochemistry and Photobiology*, *41*(4), 481–486.

Bergelson, J., Stahl, E., Dudek, S., & Kreitman, M. (1998). Genetic Variation Within and Among Populations of *Arabidopsis thaliana*. *Genetics*, *148*(3), 1311–1323.

Betts, M., & Russell, R. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for Geneticists, M.R. Barnes, I.C. Gray eds, Wiley*.

Blanc, G., Hokamp, K., & Wolfe, K. (2003). A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the *Arabidopsis* Genome. *Genome Research*, *13*, 137–144. doi:10.1101/gr.751803

Bloomer, R., Juenger, T., & Symonds, V. (2012). Natural variation in *GL1* and its effects on trichome density in *Arabidopsis thaliana*. *Molecular Ecology*, *21*(14), 3501–3515. doi:10.1111/j.1365-294X.2012.05630.x

Bongue-Bartelsman, M., & Phillips, D. (1995). Nitrogen stress regulates gene expression of enzymes in the flavonoid biosynthetic pathway of tomato. *Plant Physiology and Biochemistry*, *33*(5), 539–546.

Borevitz, J., Xia, Y., Blount, J., Dixon, R., & Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylproanoid biosynthesis. *Plant Cell*, *12*, 2383–2394.

Bradbury, P., Zhang, Z., Kroon, D., Casstevens, T., Ramdoss, Y., & Buckler, E. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635.

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, *268*, 78–94.

Burger, J., & Edwards, G. (1996). Photosynthetic efficiency, and photodamage by UV and visible radiation, in red versus green leaf coleus varieties. *Plant and Cell Physiology*, *37*(3), 395–399.

Cannon, S., Mitra, A., Baumgarten, A., Young, N., & May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology*, *4*(10).

Charlesworth, D., & Vekemans, X. (2005). How and when did *Arabidopsis thaliana* become highly self-fertilising. *BioEssays*, *27*(5), 472–476.

Cheung, V., Conlin, L., Weber, T., Arcaro, M., Jen, K., Morley, M., & Spielman, R. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, *33*, 422–425.

Close, D., & Beadle, C. (2003). The ecophysiology of foliar anthocyanin. *Botanical Review*, *69*(2), 149–161.

Cominelli, E., Gusmaroli, G., Allegra, D., Galbiati, M., Wade, H., Jenkins, G., & Tonelli, C. (2008). Expression analysis of anthocyanin regulatory genes in response to different light qualities in *Arabidopsis thaliana*. *Journal of Plant Physiology*, *165*(8), 886–894.

Costa-Arbulu, C., Gianoli, E., Gonzalez, W., & Neimeyer, H. (2001). Feeding by the Aphid *Siphia flava* produces a reddish spot on leaves of *Sarghum halepense*: an induced defense? *Journal of Chemical Ecology*, *27*(2), 273–283.

Cubas, P., Vincent, C., & Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, *401*, 157–161.

Davis, B., Poon, A., & Whitlock, M. (2009). Compensatory mutations are repeatable and clustered within proteins. *Proceedings of the Royal Society Biological Sciences*, *276*, 1823–1827. doi:10.1098/rspb.2008.1846

De Bono, M., & Bargmann, C. (1998). Natural Variation in a Neuropeptide Y Receptor Homolog Modifies Social Behavior and Food Response in *C. elegans*. *Cell*, *94*, 679–689.

De Pascual-Teresa, S., Moreno, D., & Garcia-Viguera, C. (2010). Flavanols and Anthocyanins in Cardiovascular Health: A Review of Current Evidence. *International Journal of Molecular Sciences*, *11*(4), 1679–1703.

Deikman, J., & Hammer, P. (1995). Induction of anthocyanin accumulation by cytokinins in *Arabidopsis thaliana*. *Plant Physiology*, *108*(1), 47–57.

Doyle, J., & Doyle, J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical bulletin*, *19*, 11–15.

Drumm-Herrel, H., & Mohr, H. (2006). Photosensitivity of seedlings differing in their potential to synthesize anthocyanin. *Physiologia Plantarum*, *64*(1), 60–66.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., *et al*. (2010). *Geneious*. Retrieved from http://www.geneious.com

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., & Lepiniec, L. (2010). MYB transcription factors in *Arabidopsis*. *Trends in Plant Science*, *15*(10), 1360–1385. doi:10.1016/j.tplants.2010.06.005

Feyissa, D., Lovdal, T., Olsen, K., Slimestad, R., & Lillo, C. (2009). The endogenous *GL3*, but not *EGL3*, gene is necessary for anthocyanin accumulation as induced by nitrogen depletion in *Arabidopsis* rosette stage leaves. *Planta*, *230*(4), 747–754.

Focks, N., Sagasser, M., Weisshaar, B., & Benning, C. (1999). Characterization of *tt15*, a novel transparent testa mutation of *Arabidopsis thaliana* (l.) Heynh. *Planta*, *208*, 352–357.

Gepstein, S., Sabehi, G., Carp, M., Hajouj, T., Nesher, M., *et al*. (2003). Large-scale identification of leaf senescence-associated genes. *The Plant Journal*, *36*, 629–642.

Gonzalez, A., Mendenhall, J., Huo, Y., & Lloyd, A. (2009). TTG1 complex MYBs, *MYB5* and *TT2*, control outer seed coat differentiation. *Developmental Biology*, *325*(2), 412–421.

Gonzalez, A., Zhao, M., Leavitt, J., & Lloyd, A. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. *The Plant Journal*, *53*, 814–827.

Gould, K., Markham, K., Smith, R., & Goris, J. (2000). Functional role of anthocyanins in the leaves of *Quintinia serrata* A. Cunn. *Journal of Experimental Biology*, *51*(347), 1107–1115.

Gould, K., McKelvie, J., & Markham, K. (2002). Do anthocyanins function as antioxidants in leaves? Imaging of H_2O_2 in red and green leaves after mechanical injury. *Plant, Cell and Environment*, *25*(10), 1261–1269.

Hauser, M., Harr, B., & Schlotterer, C. (2001). Trichome Distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: Molecular Analysis of the Candidate Gene *GLABROUS1*. *Molecular Biology and Evolution*, *18*(9), 1754–1763.

Hered, J. (1989). A Genetic and Phenotypic Description of *Eceriferum* (*cer*) Mutants in *Arabidopsis thaliana*. *Journal of Heredity*, *80*(2), 118–122.

Hilscher, J., Schlotterer, C., & Hauser, M. (2009). A single amino acid replacement in *ETC2* shapes trichome patterning in natural *Arabidopsis* populations. *Current Biology*, *19*, 1747–1751.

Hoch, W., Zeldin, E., & McCown, B. (2001). Physiological significance of anthocyanins during autumnal leaf senescence. *Tree Physiology*, *21*(1), 1–8.

Holton, T., & Cornish, E. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. *American Society of Plant Physiologists*, *7*(7), 1071–1083.

Innan, H., Tajima, F., Terauchi, R., & Miyashita, N. (1996). Intragenic recombination in the *ADH* locus of the wild plant *Arabidopsis thaliana*. *Genetics*, *143*, 1761–1770.

Irish, V., & Litt, A. (2005). Flower development and evolution: gene duplication, diversification and redeployment. *Current Opinion in Genetics & Development*, *15*(4), 454–460.

Jakopic, J., Stampar, F., & Veberic, R. (2009). The influence of exposure to light on the phenolic content of "Fuji" apple. *Scientia Horticulturae*, *123*(2), 234–239.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., & Dean, C. (2000). Molecular Analysis of *FRIGIDA*, a Major Determinant of Natural Variation in *Arabidopsis* Flowering Time. *Science*, *290*(5490), 344–347.

Kang, Y., Kirik, V., Hulskamp, M., Nam, K., Hagely, K., Lee, M., & Schiefelbein, J. (2009). The *MYB23* Gene Provides a Positive Feedback Loop for Cell Fate Specification in the *Arabidopsis* Root Epidermis. *The Plant Cell*, *21*, 1080–1094.

Kawabe, A., Innan, H., Terauchi, R., & Miyashita, N. (1997). Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. *Molecular Biology and Evolution*, *14*, 1303–1315.

Kim, S., Plagnol, V., Hu, T., Toomajian, C., Clark, R., *et al*. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, *39*, 1151–1155.

Kliebenstein, D., Kroymann, J., Brown, P., Figuth, A., Pederson, D., Gerhenzon, J., & Mitchell-Olds, T. (2001). Genetic Control of Natural Variation in *Arabidopsis* Glucosinolate Accumulation. *Plant Physiology*, *126*(2), 811–825.

Koch, M., Haubold, B., & Mitchell-Olds, T. (2001). Molecular systematics of the Brassicaceae: Evidence from coding plastidic *MATK* and nuclear *CHS* sequences. *American Journal of Botany*, *88*(2), 534–544.

Kong, J., Chia, L., Goh, N., Chia, T., & Brouillard, R. (2003). Analysis and biological activities of anthocyanins. *Phytochemistry*, *64*(5), 923–933.

Koorneef, M. (1981). The complex syndrome of TTG mutants. *Arabidopsis Information Service*, *18*, 45–51.

Koorneef, M., Alonso-Blanco, C., & Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology*, *55*, 141–172.

Kover, P., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I., *et al*. (2009). A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genetics*, *5*(7), e1000551.

Kranz, H., Denekamp, M., Greco, R., Jin, H., Leyva, A., *et al*. (1998). Towards functional characterisation of the members of the R2R3-MYB gene family from *Arabidopsis thaliana*. *The Plant Journal*, *16*(2), 263–276.

Kursar, T., & Coley, P. (1992). Delayed development of the phytosynthetic apparatus in tropical rain forest species. *Functional Ecology*, *6*(4), 411–422.

Lacampagne, S., Gagne, S., & Geny, L. (2010). Involvement of Abscisic Acid in Controlling the Proanthocyanidin Biosynthesis Pathway in Grape Skin: New Elements Regarding the Regulation of Tannin Composition and Leucoanthocyanidin Reductase (LAR) and

Anthocyanidin Reductase (ANR) Activities and Expression. *Journal of Plant Growth Regulation*, *29*(1), 81–90.

Lamesch, P., Berardini, T., Li, D., Swarbeck, D., Wilks, C., *et al*. (2010). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *TAIR*. Retrieved from www.arabidopsis.org

Lea, U., Slimestad, R., Smedvig, P., & Lillo, C. (2007). Nitrogen deficiency enhances expression of specific MYB and bHLH transcription factors and accumulation of end products in the flavonoid pathway. *Planta*, *225*, 1245–1253.

Lee, M., & Schiefelbein, J. (2001). Developmentally distinct MYB genes encode functionally equivalent proteins in *Arabidopsis*. *Development*, *128*, 1539–1546.

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics*, *20*(3), 116–122.

Leitch, I., & Bennett, M. (1997). Polyploidy in angiosperms. *Trends in Plant Science Reviews*, *2*(12), 470–476.

Li, B., Suzuki, J., & Hara, T. (1998). Latitudinal variation in plant size and relative growth rate in *Arabidopsis thaliana*. *Oecologia*, *115*, 293–301.

Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*, 1451–1452.

Lillo, C., Lea, U., & Ruoff, P. (2008). Nutrient depletion as a key factor for manipulating gene expression and product formation in different branches of the flavanoid pathway. *Plant Cell and Environment*, *31*, 587–601.

Lindoo, S., & Caldwell, M. (1978). Ultraviolet-B Radiation-induced Inhibition of Leaf Expansion and Promotion of Anthocyanin Production: Lack of Involvement of the Low Irradiance Phytochrome System. *Plant Physiology*, *61*(2), 278–282.

Longsreth, D., & Nobel, P. (1980). Nutrient Influences on Leaf Photosynthesis: Effects of Nitrogen, Phosphorus, and Potassium for *Gossypium hirsutum* L. *Plant Physiology*, *65*(3), 541–543.

Lovisolo, C., Perrone, I., Carra, A., Ferrandino, A., Flexas, J., Medrano, H., & Schubert, A. (2010). Drought-induced changes in development and function of grapevine (*Vitis* spp.) organs and in their hydraulic and non-hydraulic interactions at the whole-plant level: a physiological and molecular update. *Functional Plant Biology*, *37*(2), 98–116.

Machanick, P., & Bailey, T. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, *27*(12), 1696–1697.

Mancinelli, A., Hoff, A., & Cottrell, M. (1988). Anthocyanin Production in Chl-Rich and Chl-Poor Seedlings. *Plant Physiology*, *86*, 652–654.

Martin, C., Prescott, A., Mackay, S., Bartlett, J., & Vrijlandt, E. (1991). Control of Anthocyanin Biosynthesis in Flowers of *Antirrhinum majus*. *The Plant Journal*, *1*(1), 37–49.

Mathur, A., Mathur, A., Gangwar, A., Yadav, S., Verma, P., & Sangwan, R. (2010). Anthocyanin production in a callus line of *Panax sikkimensis* Ban. *In Vitro Cellular & Developmental Biology-Plant*, *46*(1), 13–21.

Matile, P. (2000). Biochemistry of Indian summer: physiology of autumnal leaf coloration. *Experimental Gerontology*, *35*(2), 145–158.

Mauricio, R., Stahl, E., Korves, T., Tian, D., Kreitman, M., & Bergelson, J. (2003). Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. *Genetics*, *163*, 735–746.

McKhann, H., Camilleri, C., Berard, A., Bataillon, T., David, J., *et al*. (2003). Nested core collections maximising genetic diversity in *Arabidopsis thaliana*. *The Plant Journal*, *38*, 193–202.

Mihai, R., Mitoi, M., Brezeanu, A., & Cogalniceanu, G. (2010). Two-stage system, a possible strategy for the enhancement of anthocyanin biosynthesis in a long-term grape callus culture. *Romanian Biotechnological Letters*, *15*(1), 5025–5033.

Mission, J., Raghothama, K., Jain, A., Jouhet, J., Block, M., *et al*. (2005). A genome-wide transcriptional analysis using *Arabidopsis thaliana* Affymetrix gene chips determined plant responses to phosphate deprivation. *PNAS*, *102*, 11934–11939.

Morcuende, R., Bari, R., Gibon, Y., Zheng, W., Pant, B., *et al*. (2007). Genome-wide reprogramming of metabolism and regulatory networks of *Arabidopsis* in response to phosphorus. *Plant, Cell and Environment*, *30*, 85–112.

Moreno, F., Monagas, M., Blanch, G., Bartolome, B., & del Castillo, M. (2010). Enhancement of anthocyanins and selected aroma compounds in strawberry fruits through methyl jasmonate vapor treatment. *European Food Research and Technology*, *230*(6), 989–999.

Muller, R., Morant, M., Jarmer, H., Nilsson, L., & Nielson, T. (2007). Genome-wide analysis of the *Arabidopsis* leaf transcriptome reveals interaction of phosphate and sugar metabolism. *Plant Physiology*, *143*, 156–171.

Niu, S., Xu, C., Zhang, W., Zhang, B., Li, X., *et al*. (2010). Coordinated regulation of anthocyanin biosynthesis in Chinese bayberry (*Myrica rubra*) fruit by a R2R3 MYB transcription factor. *Planta*, *231*(4), 887–899.

Oda, M., Furukawa, K., Ogata, K., Sarai, A., Ishii, S., Nishimura, Y., & Nakamura, H. (1997). Identification of indispensable residues for specific DNA-binding in the imperfect tandem repeats of c-MYB R2R3. *Protein engineering design & selection*, *10*(12), 1407–1414. doi:10.1093/protein/10.12.1407

Ogata, K., Kanei-Ishii, C., Sasaki, M., Hatanaka, H., Nagadoi, A., *et al*. (1996). The cavity in the hydrophobic core of MYB DNA-binding domain is reserved for DNA recognition and trans-activation. *Nature Structural Biology*, *3*(2), 178–187.

Ohno, S. (1970). *Evolution by gene duplication*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Oleksiak, M., Churchill, G., & Crawford, D. (2002). Variation in gene expression within and among natural populations. *Nature Genetics*, *32*, 261–266.

Powles, S. (1984). Photoinhibition of photosynthesis induced by visible light. *Review of Plant Physiology*, *35*(1), 15–44.

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., & Van de Peer, Y. (2003). Investigating ancient duplication events in the *Arabidopsis* genome. *Journal of Structural and Functional Genomics*, *3*, 117–129.

Rajagopalan, R., Vaucheret, H., Trejo, J., & Bartel, D. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Development*, *20*, 3407–3425.

Ramsay, N., & Glover, B. (2005). MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends in Plant Science*, *10*(2), 63–70.

Riechmann, J., Heard, J., Martin, G., Reuber, L., Jiang, C., *et al*. (2000). *Arabidopsis* Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. *Science*, *290*, 2105–2110.

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

Rose, L., Bittner-Eddy, P., Langley, C., Holub, E., Michelmore, E., & Beynon, J. (2004). The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. *Genetics*, *166*, 1517–1527.

Rowan, D., Cao, M., Kui, L., Cooney, J., Jensen, D., *et al*. (2009). Environmental regulation of leaf colour in red 35S:*PAP1 Arabidopsis thaliana*. *New Phytologist*, *182*, 102–115.

Rozen, S., & Skaletsky, H. (2000). *Primer3 on the WWW for general users and for biologist programmers*. Retrieved from http://fokker.wi.mit.edu/primer3/

Sarma, A., & Sharma, R. (1999). Anthocyanin-DNA copigmentation complex: mutual protection against oxidative damage. *Phytochemistry*, *52*(7), 1313–1318.

Scmuths, H., Hoffman, M., & Bachmann, K. (2004). Geographic Distribution and Recombination of Genomic Fragments on the Short Arm of Chromosome 2 of *Arabidopsis thaliana*. *Plant Biology*, *6*(2), 128–139.

Sharbel, T., Haubold, B., & Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology*, *9*, 2109–2118.

Shi, M., & Xie, D. (2010). Features of anthocyanin biosynthesis in *pap1-D* and wild-type *Arabidopsis thaliana* plants grown in different light intensity and culture media conditions. *Planta*, *231*, 1385–1400.

Shiu, S., Karlowski, W., & Pan, R. (2004). Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell*, *16*, 1220–1234.

Simillion, C., Vandepoele, K., Van Montagu, M., Zabeau, M., & Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. *PNAS*, *99*(21), 13627–13632.

Simon, M., Simon, A., Martins, F., Botran, L., Tisne, S., *et al*. (2012). DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. *The Plant Journal*, *69*, 1094–1101.

Solfanelli, C., Poggi, A., Loreti, E., Alpi, A., & Perata, P. (2006). Sucrose-specific Induction of the Anthocyanin Biosynthetic Pathway in *Arabidopsis*. *Plant Physiology*, *140*, 637–646.

Springbob, K., Nakajima, J., Yamazaki, M., & Saito, K. (2002). Recent Advances in the Biosynthesis and Accumulation of Anthocyanins. *Natural Products Reports*, *20*, 288–303.

Steyn, W., Wand, S., Holcroft, D., & Jacobs, G. (2002). Anthocyanins in vegetative tissues: a proposed unified function in photoreception. *New Phytologist*, *155*(3), 349–361.

Stintzing, F., & Carle, R. (2004). Functional properties of anthocyanins and betalains in plants, food and in human nutrition. *Trends in Food Science and Technology*, *15*(1), 19–38.

Stracke, R., Werber, M., & Weisshaar, B. (2001). The R2R3-MYB gene family in *Arabidopsis thaliana*. *Current Opinion in Plant biology*, *4*, 447–456.

Su, V., & Hsu, B. (2010). Transient Expression of the Cytochrome p450 *CYP78A2* Enhances Anthocyanin Production in Flowers. *Plant Molecular Biology Reporter*, *28*(2), 302–308.

Symonds, V., Hatlestad, G., & Lloyd, A. (2011). Natural allelic variation denies a role for *ATMYBC1*: trichome cell fate determination. *PLoS Genetics*, *7*.

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

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*, 796–815.

Thiele, A., Krause, G., & Winter, K. (1998). In Situ study of photoinhibition of photosynthesis and xanthophyll cycle activity in plants growing in natural gaps of the tropical forest. *Australian Journal of Plant Physiology*, *25*(2), 189–196.

Tian, D., Araki, H., Stahl, E., Bergelson, J., & Kreitman, M. (2002). Signature of balancing selection in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 11525–11530.

Tucic, B., Vuleta, A., & Jovanovic, S. (2009). Protective Function of Foliar Anthocyanins: In Situ Experiments on a Sun-Exposed Population of *Iris pumila* L. (Iridaceae). *Polish Journal of Ecology*, *57*(4), 779–783.

Vanderauwera, S., Zimmerman, P., Rombauts, S., Vandenabeele, S., *et al*. (2005). Genome-Wide Analysis of Hydrogen Peroxide-Regulated Gene Expression in Arabidopsis Reveal a High Light-Induced Transcriptional Cluster Involved in Anthocyanin Biosynthesis. *Plant Physiology*, *139*, 806–821.

Velten, J., Cakir, C., & Cazzonelli, C. (2010). A spontaneous dominant-negative mutation within a 35S::*AtMYB90* transgene inhibits flower pigment production in tobacco. *PLoS ONE*, *5*(3), e9917.

Vollmannova, A., Toth, T., Urminska, D., Polakova, Z., Timoracka, M., & Margitanova, E. (2009). Anthocyanin Content in Blueberries (*Vaccinium corymbosum* L.) in Relation to Freezing Duration. *Czech Journal of Food Sciences*, *27*, 5204–5206.

Vuleta, A., Jovanovic, S., Seslija, D., & Tucic, B. (2010). Seasonal dynamics of foliar antioxidative enzymes and total anthocyanins in natural populations of *Iris pumila* L. *Journal of Plant Ecology-UK*, *3*(1), 59–69.

Wang, B., He, R., & Li, Z. (2010). The Stability and Antioxidant Activity of Anthocyanins from Blueberry. *Food Technology and Biotechnology*, *48*(1), 42–49.

Wang, H., Cao, G., & Prior, R. (1997). Oxygen Radical Absorbing Capacity of Anthocyanins. *Journal of Agricultural and Food Chemistry*, *45*(2), 304–309.

Xue, W., Xing, Y., Weng, X., Zhao, X., Tang, W., *et al.* (2008). Natural Variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nature Genetics*, *40*, 761–767.

Yang, Y., Lai, K., Tai, P., & Li, W. (1999). Rates of Nucleotide Substitution in Angiosperm Mitochondrial DNA Sequences and Dates of Divergence Between Brassica and Other Angiosperm Lineages. *Journal of Molecular Evolution*, *48*, 597–604.

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., *et al*. (2006). The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Molecular Biology*, *60*, 107–124. doi:10.1007/s11103-005-2910-y

Yuan, Y., Chiu, L., & Li, L. (2009). Transcriptional regulation of anthocyanin biosynthesis in red cabbage. *Planta*, *230*(6), 1141–1153.

Zhang, F., Gonzalez, A., Zhao, M., Payne, C., & Lloyd, A. (2003). A network of redundant bHLH proteins functions in all TTG1-dependent pathways of *Arabidopsis*. *Development*, *130*, 4859–4869.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol*, *7*(1-2), 203–214.

Zimmermann, I., Heim, M. A., Weisshaar, B., & Uhrig, J. F. (2004). Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. *The Plant Journal*, *40*(1), 22–34.

Ziolkowski, P., Kaczmarek, M., Babula, D., & Sadowski, J. (2006). Genome evolution in *Arabidopsis*/Brassica: conservation and divergence of ancient rearranged segments and their breakpoints. *Plant Journal*, *47*(1), 63–74.

7. Appendix 1-The Versailles Core Collection of Natural Accessions of *Arabidopsis thaliana*

Accession Name	Versailles ID	Location
Akita	252AV	Akita prefecture, Japan
Alc-0	178AV	Alcalá de Henares, Spain
Bl-1	42AV	Bologna, Italy
Bla-1	76AV	Gerona, Spain
Blh-1	180AV	Bulhary, Czech Republic
Bur-0	172AV	Burren, Ireland
Can-0	163AV	Canary Islands
Ct-1	162AV	Catania, Italy
Cvi-0	166AV	Cape Verdi Islands
Edi-0	83AV	Edinburgh, Scotland
Enkheim-T	197AV	Enkheim, Germany
Ge-0	101AV	Geneva, Switzerland
Gre-0	200AV	Greenville, Michigan, USA
Ishikawa	253AV	Uchinada, Japan
Ita-0	157AV	Tazekka, Morocco
Jea	25AV	St Jean Cap Ferret, France
Jm-0	206AV	Jamolice, Czech Republic
Kn-0	70AV	Kaunas, Lithuania
Kondara	190AV	Kondara, Tadjikistan
Lip-0	63AV	Chrzanow, Poland
Mh-1	215AV	Muhlen, Poland
Ms-0	93AV	Moscow, Russia
Mt-0	94AV	Martuba, Libya
N13	266AV	Konchezero, Russia
N14	267AV	Sampo Hill, Russia
N6	262AV	Karelia, Russia
N7	263AV	Pinguba Bay, Russia
Nok-1	95AV	Noordwijk, Netherlands
Oy-0	224AV	Oystese, Norway
Pa-1	50AV	Palermo, Italy
Pi-0	40AV	Pitzal/Tirol, Western Austria
Pyl-1	8AV	Le Pyla, France
Ran	21AV	Rance Estuary, France
Ri-0	160AV	Richmond, B.C, Canada
Rld-2	229AV	Rzhev, Russia
Rubezhnoe-1	231AV	Rubezhnoe, Ukraine
Sah-0	233AV	Alhambra, Spain
Sakata	257AV	Sakata, Japan
Sap-0	234AV	Slapy, Czech Republic
Sav-0	235AV	Slavice, Czech Republic
Shahdara	236AV	Pamiro-Alaya, Tajikistan
Sp-0	53AV	Berlin, Germany
St-0	62AV	Stockholm, Sweden
Stw-0	92AV	Orel, Russia
Ta-0	56AV	Tabor, Czech Republic
Te-0	68AV	Tenala, Finland
Tsu-0	91AV	Tsu, Japan
Yo-0	250AV	Yosemite Nat. Park, USA

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

CERTIFICATE OF REGULATORY COMPLIANCE

This is to certify that the research carried out in the Masterate Thesis entitled Natural variation in the serially duplicated production of anthocyanin pigment loci and anthocyanin accumulation in Arabidopsis thaliana (Brassicaceae) in the Institute of Fundamental Sciences at Massey University, New Zealand:

(a) is the original work of the candidate, except as indicated by appropriate attribution in the text and/or in the acknowledgements;

(b) that the text, excluding appendices/annexes, does not exceed 40,000 words;

(c) all the ethical requirements applicable to this study have been complied with as required by Massey University, other organisations and/or committees _____ which had a particular association with this study, and relevant legislation.

Please insert Ethical Authorisation code(s) here: (if applicable) _____

_____

Candidate's Name: Matt Butcher

Signature: 

Date: 12/6/2013

Supervisor's Name: Dr. Vaughan Symonds

Signature: _____

Date: 06 DEC 2013

Institute of Fundamental Sciences
Private Bag 11222, Palmerston North 4442, New Zealand. **T** +64 6 356 9099 **F** +64 6 350 5682 **E** ifs@massey.ac.nz **http://ifs.massey.ac.nz**