

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

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

Acknowledgements	iii
Contents	iv
Figures	vii
Tables.....	xii
1. Abstract.....	1
2. Introduction	2
2.1 The Biological Roles of Anthocyanins.....	3
2.2 Anthocyanin Biosynthesis and Regulation of Production.....	8
3. Molecular Analysis of the <i>PAP</i> Genes.....	15
3.1 Introduction	15
3.2 Materials and Methods.....	18
3.2.1 Plant Materials.....	18
3.2.2 <i>PAP</i> and <i>WER</i> sequencing.....	19
3.2.3 Gene Cloning.....	24
3.2.4 Molecular data analysis.....	25
3.3 Results.....	28
3.3.1 Nucleotide diversity and patterns of polymorphism in genomic alignments of the <i>PAP</i> and <i>WER</i> loci.....	28
3.3.1.1 <i>PAP1</i>	28
3.3.1.1.1 <i>PAP1</i> haplogroup A.....	28
3.3.1.1.2 <i>PAP1</i> haplogroup B.....	31
3.3.1.2 <i>PAP2</i>	32
3.3.1.3 <i>PAP3</i>	33
3.3.1.4 <i>PAP4</i>	34
3.3.1.5 <i>WER</i>	36
3.3.2 Intragenic variation of the coding regions	36
3.3.2.1 <i>PAP1</i>	36
3.3.2.2 <i>PAP2</i>	42
3.3.2.3 <i>PAP3</i>	44
3.3.2.4 <i>PAP4</i>	47
3.3.2.5 <i>WER</i>	49
3.3.2.6 R2R3 MYB regions.....	51

3.3.2.6.1	52
3.3.2.6.2	53
3.3.2.6.3	54
3.3.2.6.4	55
3.3.2.6.5	55
3.3.3 Intergenic molecular evolution amongst the PAPs	56
3.3.3.1 PAP1 and PAP2.....	57
3.3.3.2 PAP1 and PAP3.....	59
3.3.3.3 PAP1 and PAP4.....	61
3.3.3.4 PAP2 and PAP3.....	63
3.3.3.5 PAP2 and PAP4.....	65
3.3.3.6 PAP3 and PAP4.....	67
3.3.4 Analysing the phylogenetic relationships of the PAP genes	69
3.3.5 Linkage disequilibrium of the PAP genes	72
3.3.6 Unique motifs identifying the PAP genes	74
3.3.6.1 Motifs in the R2R3-MYB region.....	74
3.3.6.2 Motifs in the undefined region.	76
3.3.6.3 De novo motif identification.	77
3.3.7 The PAP genes and transcriptional regulation	78
3.4 Discussion	79
3.4.1 Variation and selection between the PAP genes.....	79
3.4.2 Variation and selection within the PAP genes.....	80
3.4.3 Mutations affecting the MYB domains.....	82
3.4.4 Phylogenetic relationships between the PAP genes	82
3.4.5 Allele association between the PAP genes.....	84
3.4.6 Identifying MYB genes using motifs	85
3.4.7 Biallelic patterns of the PAP genes.....	86
4. An Investigation of the Genetic Architecture of Anthocyanin Accumulation.....	88
4.1 Introduction	88
4.2 Materials and Methods.....	89
4.2.1 Plant material and growth conditions.....	89
4.2.2 Pigment extraction and analysis.	90
4.3 Results.....	91
4.3.1 Heritability and mapping of anthocyanin accumulation	91

4.4 Discussion	97
5. Conclusion.....	99
6. References Cited	103
7. Appendix 1-The Versailles Core Collection of Natural Accessions of <i>Arabidopsis thaliana</i>	117

Figures

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 3 β -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). 9

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. 13

Figure 4 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..... 29

Figure 5 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..... 30

Figure 6 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..... 31

Figure 7 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..... 32

Figure 8 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....**Error! Bookmark not defined.**

Figure 9 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..... 35

Figure 10 Sliding window analysis of nucleotide diversity (π) for an alignment of *WER* genomic sequences from 48 accessions of *Arabidopsis thaliana* with π 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..... 36

Figure 11 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. 38

Figure 12 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. 42

Figure 13 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. 43

Figure 14 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). 44

Figure 15 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..... 44

Figure 16 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. 46

Figure 17 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..... 47

Figure 18 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. 48

Figure 19 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. 50

Figure 20 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). 51

Figure 21 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..... 57

Figure 22 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. 59

Figure 23 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. 61

Figure 24 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. 63

Figure 25 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. 65

Figure 26 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. 67

Figure 27 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. 69

Figure 28 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..... 70

Figure 29 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. 71

Figure 30 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. 72

Figure 31 Linkage disequilibrium analysis of mutations of the *PAP* genes. Intragenic measures of linkage disequilibrium are shown boxed. The extent of linkage disequilibrium (R^2) above the black diagonal line. 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. 74

Figure 32 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. 75

Figure 33 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. 76

Figure 34 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. 92

Figure 35 Chromosome maps of *Arabidopsis thaliana* with associated loci plotted against logP scores. Each point on the plot represents 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). 93

Tables

Table 1 Gene Nomenclature	13
Table 2 List of primers used in PCR and sequencing reactions	20
Table 3 Standard Polymerase Chain Reaction protocols.....	22
Table 4 Standard sequencing reaction protocols.....	23
Table 5 <i>PAP</i> amino acid replacements and indels identified from 48 <i>Arabidopsis thaliana</i> accessions	39
Table 6 A summary of measures of nucleotide diversity across the genomic and inferred coding sequences of the <i>PAP</i> and <i>WER</i> genes	42
Table 7 Summary of <i>Ka/Ks</i> averages of different gene regions of the <i>PAP</i> genes.....	58
Table 8 Summary of the location and function of genes likely underlying loci associated with anthocyanin accumulation. The '/' between genes indicates that.....	94
Table 9 Location and function of genes involved in regulation of anthocyanin accumulation and biosynthesis	95
Table 10 Location and function of senescence-associated genes	96

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 biosynthesis 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.