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

Gene duplication fate in a genetic pathway context: An examination of the trichome initiation pathway in the allopolyploid genus *Pachycladon* (Brassicaceae)

A thesis presented in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Plant Biology

At Massey University,

Manawatu,

New Zealand

Thilini Madara Warusawithana

2024

Abstract

The evolutionary significance of polyploidy is difficult to overstate and one of the primary foci of modern polyploidy research is on the fates of duplicated genes (homeologs). The impact of gene duplication on biological evolution is strongly related to evolutionary novelty, genome architecture, and even speciation. However, current approaches to study homeolog fates are largely descriptive (e.g., whole genome scans) and lack a critical context: the genetic pathway. As a result, relatively little is known about how homeologs interactively behave within a duplicated genetic pathway to form protein-protein interactions and, how homeolog functional fates are associated with particular gene attributes such as pleiotropy and epistasis. We examined the well-characterized trichome initiation pathway in the allopolyploid *Pachycladon* (Brassicaceae) in this study to investigate how integrated genetic systems respond to the common phenomenon of duplication. The genus is the product of a hybridization event that took place during the Pleistocene ~0.8-1.6 mya, between two highly diverged lineages, followed by polyploidy and diversification. Considering morphological and phylogenetic diversity, four *Pachycladon* species, *P. enysii*, *P. fastigiatum*, *P. novae-zealandiae* and *P. cheesemanii* were selected to study the trichome initiation pathway focusing on the key genes *TTG1, GL1*, *GL3,* and *EGL3*.

Pachycladon's parental origins are not known exactly yet, despite that previous studies describe that one of the parental genomes is closely related to *A. thaliana*. Hence, as the first step, we set out to isolate the duplicated copies (homeologs) of the four genes of interest from each of the four study species. Those sequences were then analyzed to identify the progenitor origins of each homeolog for each gene to 'assemble' the progenitor pathways. The gene trees resulting from phylogenetic analyses were consistent in the placements of homeologs for each gene, positioning one of the copies (Type 1) in the *Camelineae* lineage (lineage I) and other copy (Type 2) outside of (and typically sister to) Lineage I.

Categorization of the homeologs of each gene into progenitor Types facilitated the analysis of sequence divergence and expression studies to provide more insights into the

molecular evolution of the pathway. The Type 2 copies of all genes were found to have greater variation than the Type 1 copies. Despite several mutations with likely functional effects being identified, Ka/Ks analyses suggested that both copies of most genes in most species have been under the influence of purifying selection. qPCR analyses of the relative expression of gene homeologs suggested similar gene expression patterns in four *Pachycladon* species and across two sample types, whole seedlings, and mature leaves. Based on the results, we assumed multiple molecular mechanisms overlap controlling the expression of the homeologs such as relative gene dosage balance, subgenome dominance, and pleiotropy. The most general finding was a trend showing preferential expression of Type 1 homeologs observed across genes and species (although there were exceptions to this pattern). Based on other studies, the Type 1 homeologs are likely to be of paternal origin, indicating a paternal bias in homeolog expression. In contrast to the other genes examined, the highly pleiotropic gene, *TTG1*, was found to have very similar levels of homeolog expression across all species. These results provide a detailed analysis of the molecular evolution of a genetic pathway following duplication via allopolyploidy.

Acknowledgments

My PhD research was the most challenging period in my life, so I couldn't accomplish it without the invaluable support of many people around me. I would first like to express my sincere gratitude to my supervisor, Dr. Vaughan Symonds, for offering me the opportunity to start the project and supporting me in securing the Massey Doctoral Scholarship, guiding me throughout with his expertise knowledge, and patience in undertaking the research and writing of this thesis. I am also grateful to my co-supervisor Dr. Matthieu Vignes for the insightful comments, encouragement, and guidance, and Dr. Anja Schiemann for supporting me with the expression analysis. My sincere thanks also go to the Massey Doctoral Scholarship, Heseltine Trust Scholarship, Emergency COVID-19 Hardship Grant, and Putea Tautoko-Doctoral Financial Support Grant for the funding that enabled me to undertake this research. Previous studies on *Pachycladon* by Dr. Rebecca Bloomer, and Dr. Vaughan Symonds, especially the primers designed gave me a head start to begin my research. I can't forget Assoc. Prof Jennifer Tate for the comments and encouragement given, and Assoc. Prof Catherine Whitby and all the administrative staff supported all my suspension and extension applications based on my circumstances. I am grateful to my examiners Prof Mary Morgan-Richards, Dr Rebecca Bloomer, and Dr Antonio Gonzalez, and convenor Prof Barry Borman for their valuable insightful suggestions and comments on the final emendations of this thesis. Also, my sincere thanks to the lab technician, Dr. Prashant Joshi, for his help and knowledge in the lab. I would like to thank my amazing 'LOST' lab group, especially my fellow lab mates Sidra, Weixuan, and Kay for their support and willingness to help me.

My journey was special because I had my first child born after a short time I started my study. So, my heartfelt gratitude goes to my husband Dim for his endurance and unwavering support given throughout the difficult times. I also want to thank my parents and my parents-in-law for visiting us from Sri Lanka and taking care of our son. My father has been my inspiration from childhood to explore science and climb up the ladder to become an academic expert. I cannot forget my loving son Thinon, who was beside me all the time.

Last but not least, I want to thank my managers Sanjay and Amritha at Biotelliga (Pvt) Company, for supporting my thesis writing after I started working full-time in the company. My work colleagues Kim, Mikhail, Libby, Sunita, Teck-Nam, and Renjitha encouraged me often when I was struggling with my writing with the full-time job and caring for my son and family at the same time. I would not be able to see the end of this difficult chapter in my life studying a PhD while working full-time and taking care of a child without the support from these wonderful people.

Table of Contents

Abstract ii

Acknowledgments iv

List of Figures x

List of Tables xii

List of Appendices xiii

CHAPTER ONE 1

INTRODUCTION 1

1.1 Gene duplication 1

1.1.1 Polyploidy 2
1.1.2 Origins of polyploids 4
1.1.3 Polyploidy and plant evolution – Molecular consequences of polyploidy 5
1.1.3.1 Chromosomal loss, fusion and rearrangements 5
1.1.3.2 Duplicate gene loss and retention 6
1.1.3.3 Functional divergence of retained duplicated genes 7
1.1.3.4 Differential gene expression 8
1.1.3.5 Diploidization 10
1.1.4 Ancestral polyploidization events related to the Brassicaceae lineage 11
1.1.5 Homeolog fate in a genetic pathway context 12

1.2 Brassicaceae epidermal cell fate pathway 13

1.2.1 Cell differentiation in the leaf epidermis 16
1.2.1.1 Plant trichomes and the origin of *Arabidopsis* trichomes 16
1.2.1.2 Trichome initiation pathway in *Arabidopsis* 17

1.3 *Pachycladon* (New Zealand rockcress) as a study system 18

1.3.1 Basic botany and distribution of *Pachycladon* 20
1.3.2 Allopolyploid origin of *Pachycladon* within the *Brassicaceae* phylogeny 21

1.4 Sequencing approaches to generate gene pathway data 24

1.4.1 First Generation Sequencing - Sanger Sequencing 24
1.4.2 Next Generation Sequencing (NGS) Technology 25

1.5 Gene expression analysis 27

1.5.1 qRT-PCR (Quantitative Reverse Transcription-PCR) 27

1.6 Thesis structure and objectives 30

Chapter 2 31
1.6.1 Reconstruct and characterize the “ancestral” epidermal cell fate pathways from multiple *Pachycladon* species, with a focus on the trichome initiation pathway 31
1.6.1.1 Isolate and sequence both homeologs (where present) of the primary genes that direct epidermal cell fate 31
1.6.1.2 Reconstruct and contrast the two progenitor trichome development pathways by phylogenetic analysis 32

Chapter 3 32
1.6.2 Analyzing and characterizing the molecular evolution of the trichome initiation pathway across species 33
Chapter 4 33
1.6.3 Expression analysis of homeologous genes to analyze differential expression 33
Chapter 5 34

REFERENCES 35

A1. APPENDIX 46

2. CHAPTER TWO 48

Reconstructing and characterizing *Pachycladon* "ancestral" epidermal cell fate pathways with a focus on the trichome initiation pathway 48

2.1 ABSTRACT 48

2.2 INTRODUCTION 49

2.2.1 Polyploidy and its consequences 49
2.2.2 *Pachycladon* as a study system 50
2.2.2.1 Allopolyploid origin and evolution 50
2.2.2.2 High rate of species adaptive radiation and karyotype stasis 51
2.2.2.3 Origin in New Zealand 54
2.2.3 Reconstructing *Pachycladon* "ancestral" trichome initiation pathways 55

2.3 MATERIALS AND METHODS 56

2.3.1 Plant material 56
2.3.2 Genomic DNA extraction 56
2.3.3 Sequencing methods used to generate gene sequences 57
2.3.4 Genomic DNA Sanger sequencing 58
2.3.4.1 Homeolog-specific primer design 59
2.3.4.2 Polymerase chain reaction (PCR) 61
2.3.4.3 Sanger sequencing of the homeologs from genomic PCR products from *TTG1*, *GL1*, *GL3* and *EGL3* genes from *P. enysii*, *P. fastigiatum* and *P. novae-zealandiae* 63
2.3.5 Illumina MiSeq sequencing of total genomic DNA belonging to *P. enysii*, *P. fastigiatum* and *P. novae-zealandiae* 63
2.3.6 RNA extraction, cDNA synthesis, cDNA Sanger sequencing, and cDNA MiSeq sequencing 64
2.3.7 Sequence analysis 65
2.3.7.1 Assembly of reads 65
2.3.7.2 Mapping MiSeq reads to the *Arabidopsis thaliana* homolog 66
2.3.7.3 Phasing homeologs 66
2.3.8 Phylogenetic analysis 69

2.4 RESULTS 71

2.4.1 Homeolog sequencing of *TTG1*, *GL1*, *GL3*, and *EGL3* from *P. enysii*, *P. fastigiatum* and *P. novae-zealandiae* 71
2.4.1.1 Sanger sequencing 71
2.4.1.2 Illumina MiSeq sequencing 71
2.4.1.3 Sequence analyses 71

2.4.2 Phylogenetic analysis of the genes to reconstruct progenitor pathways 74

2.5 DISCUSSION 81

2.5.1 Sequence analyses 81
2.5.2 Phylogenetic analysis of the genes to reconstruct progenitor pathways 83

2.6 CONCLUSION 85

A2. APPENDIX 90

3. CHAPTER THREE 92

Analyzing and characterizing the molecular evolution of the trichome initiation pathway across species 92

3.1 ABSTRACT 92

3.2 INTRODUCTION 93

3.2.1 Polyploidy, doubling of genetic pathways, and subsequent evolution 93
3.2.2 The duplicated trichome initiation pathway in *Pachycladon* 93
3.2.3 The TTG1-GL1-GL3/EGL3 conserved transactivation complex 94
3.2.4 Paralogs in the transactivation complex and their functional redundancy 97
3.2.5 Functional divergence of duplicates in gene regulatory network evolution 97
3.2.6 Mutations in protein-coding regions 99
3.2.7 *Pachycladon* divergence 101
3.2.8 Molecular evolution of the *Pachycladon* trichome initiation pathway 102

3.3 MATERIALS AND METHODS 103

3.3.1 Haplotype network construction 103
3.3.2 DNA polymorphism analysis 103
3.3.3 DNA divergence analysis in synonymous and nonsynonymous sites 103
3.3.4 Protein analysis 104

3.4 RESULTS 107

3.4.1 Haplotype network construction 107
3.4.2 DNA polymorphism analysis 108
3.4.3 Analysis of molecular evolution using synonymous and nonsynonymous sites 109
3.4.4 Protein analysis 111
3.4.4.1 TTG1 protein analysis of significant amino acid variations 111
3.4.4.2 GL1 proteins 3D model prediction and analysis of significant amino acid variations 115

3.5 DISCUSSION 119

3.5.1 Gene polymorphisms create two distinct haplogroups 119
3.5.2 Type 2 gene copies have more variation than Type 1 gene copies 119
3.5.3 The genes are under purifying selection 121
3.5.4 Molecular evolution of the pathway proteins 122
3.5.4.1 Evolution of TTG1 duplicated protein copies 122
3.5.4.2 Evolution of GL1 duplicated protein copies 123
3.5.4.3 Evolution of GL3 duplicated protein copies 126
3.5.4.4 Evolution of EGL3 duplicated protein copies 127
3.5.4.5 Genetic redundancy of duplicates and functional divergence 127

3.5 CONCLUSION 129

REFERENCES 130

A3. APPENDIX 136

4.0 CHAPTER FOUR 138

Expression analysis of duplicated gene copies (homeologs) in *Pachycladon* trichome initiation pathway 138

4.1 ABSTRACT 138

4.2 INTRODUCTION 139

4.2.1 Polyploidy, differential expression and evolution 139
4.2.2 Evolution of expression divergence in genetic complexes/pathways following whole genome duplication 140
4.2.2.1 Gene dosage balance hypothesis 140
4.2.2.2 Subgenome dominance 141
4.2.3 Gene expression studies of *Arabidopsis* trichome initiation pathway 142
4.2.4 Studies on duplicated trichome initiation pathway outside *Arabidopsis* 144
4.2.5 Expression divergence of *Pachycladon* duplicated trichome initiation pathway 145

4.3 MATERIALS AND METHODS 147

4.3.1 Plant materials collection, RNA extraction, and cDNA synthesis 147
4.3.2 Selection of qPCR reference genes 149
4.3.3 qPCR SYBR Green primer design 150
4.3.4 Assay validation 152
4.3.4.1 Specificity analysis 152
4.3.4.2 Efficiency analysis 152
4.3.5 qPCR experiments 154
4.3.6 Data analysis and relative quantification 156
4.3.6.2 Statistical analysis of data 157
4.3.6.3 Gene expression results validation using qRAT 157

4.4 RESULTS 159

4.4.1 qPCR assay primer efficiencies 159
4.4.2 qPCR data statistical analysis 160
4.4.2.1 Consistently different homeolog expression of *GL3* in all species and developmental stages, a species-specific homeolog expression difference in *GL1*, and a general trend of parental bias 164
4.4.2.2 Overall biased expression of Type 1 gene copies 165
4.4.2.3 Homeolog expression differences in *P. cheesemanii* seedlings vs. mature leaves 165

4.5 DISCUSSION 168

4.5.1 *TTG1* sequence diversity and homeolog expression 168
4.5.2 *GL1* sequence diversity and homeolog expression 169
4.5.3 *GL3* sequence diversity and homeolog expression 170
4.5.4 *EGL3* sequence diversity and homeolog expression 172
4.5.5 Comparison between mature leaf tissue and seedling gene expression 172
4.5.6 A role for pleiotropy in homeolog expression and retention 173
4.5.7 A general finding of biased expression of Type 1 homeologs over Type 2 174
4.5.8 Evolution of relative expression of *TTG1*, *GL1*, *GL3*, and *EGL3* in the *Pachycladon* trichome initiation complex is suggested to be maintained by combinatorial effects of subgenome dominance, biased fractionation, and relative gene dosage balance 175

4.6 CONCLUSION 177
REFERENCES 178
A4. APPENDIX 185
5.0 CHAPTER FIVE 197
Thesis summary and future directions 197
5.1 Reconstruction of the two progenitor trichome development pathways 198
5.2 Molecular evolution of the duplicated genetic pathway 199
5.3 Expression analysis of the duplicated genetic pathway 201
5.4 Future directions 204
REFERENCES 206

List of Figures

Figure 1.1: Possible pathways of polyploid formation……..5

Figure 1.2: Transcription factors and proteins involved in forming the common activation complex in different epidermal cell fate determination pathways in *Arabidopsis*……...14

Figure 1.3: A schematic diagram showing the ancient origin of *Pachycladon*, hybridization, allopolyploidization, and speciation……..19

Figure 1.4: (A) Distribution of all New Zealand *Pachycladon* species (B) NeighborNet network of inter-organism genetic distances computed from 10 single-copy nuclear genes with the POFAD algorithm……...20

Figure 1.5: Morphology and growth habit of (a) *P. cheesemanii*, (b) *P.enysii*, (c) *P.fastigiatum* and (d) *P. novae-zealandiae*……...21

Figure 1.6: Sample qPCR amplification plot-a threshold level is set where the sample produces a detectable fluorescence signal over the background noise, and CT indicates the cycle number at which this occurs……...…28

Figure 2.1: Comparative cytogenetic map of duplicated Ancestral Crucifer Karyotype (ACK) (two sets of eight ancestral chromosomes (AK1- AK8)) showing collinearity relationships with ten *Pachycladon* chromosomes (PC1 - PC10)…….....…54

Figure 2.2: Procedure used to generate *Pachycladon* homeolog coding sequences……58

Figure 2.3: The *TTG1* evolutionary history inferred by using the Maximum Likelihood method and General Time Reversible model…….……75

Figure 2.4: *GL1* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model…….……77

Figure 2.5: *GL3* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model…….…78

Figure 2.6: *EGL3* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model…….……79

Figure 3.1: Schematic representation of (a) WD40 protein, (b) R2R3 MYB, and (c) bHLH TFs……96

Figure 3.2: Modes of evolutionary changes in TF-DNA binding after genome duplication..99

Figure 3.3: The median mutational effect score of amino acid substitutions……101

Figure 3.4: Median-joining haplotype network of (a) *TTG1* and (b) *GL1* coding sequences……107

Figure 3.5: The nucleotide diversity (Pi) of *Pachycladon* Type 1 homeologs and Type 2 homeologs against the nucleotide position of (a) *TTG1* and (b) *GL1* nucleotide sequences……108

Figure 3.6: Ratio of nonsynonymous nucleotide divergence to synonymous nucleotide divergence (Ka/Ks) against the nucleotide positions of *Pachycladon* genes (a) *TTG1* and (b) *GL1*……..……110

Figure 3.7: Alignment of *A. thaliana* and *Pachycladon* species TTG1 amino acid sequences……112

Figure 3.8: Alignment of *A. thaliana* and *Pachycladon* species GL1 amino acid sequences……117

Figure 3.9: Consensus sequences and level of conservation of 125 *A. thaliana* R2 (a) and R3 (b) MYB domains, aligned with *Pachycladon GL1* Type 1 and Type 2 consensus sequences……....124

Figure 3.10: Model showing the effect of the stoichiometry of gene regulatory complexes and its impact in generating phenotypes……128

Figure 4.1: *Pachycladon cheesemanii* plants and leaves collected for RNA extraction……148

Figure 4.2: Example experimental setup (96 well plate) for *P. enysii* for *TTG1* Type 1 and Type 2 homeologs……...…155

Figure 4.3: Log 2 fold difference in mean expression of Type 1 homeolog (A) to that of the Type 2 homeolog (B) (A-B log scale) in (a) *P. en* (*P. enysii*) seedling (b) *P.ch* (*P. cheesemanii*) seedling (c) *P.ch* (*P. cheesemanii*) mature leaf, (d) *P. fs* (*P. fastigiatum*) mature leaf and (e) *P. nz* (*P. novae-zealandiae*) mature leaf…….……..…163

Figure 4.4: (a) Log 2 scale relative expression differences in type-specific homeologs of the genes in young and mature *P. cheesemanii*. (b) Log fold expression difference of Type 1 homeologs and (c) Type 2 homeologs of all genes in young *P. cheesemanii* compared to its mature stage……166

List of Tables

Table 2.1: Morphological and distributional characteristics of *Pachycladon* species…..……53

Table 2.2: *Pachycladon* plant materials collected and used……..…56

Table 2.3: Details of used primers……60

Table 2.4: Standard Phusion PCR recipe calculated for 10 µL total reaction……..…61

Table 2.5: Thermocycling conditions used for routine PCR……61

Table 2.6: A tailing with *Taq* polymerase……62

Table 2.7: Summary of all the methods utilized to generate and validate *Pachycladon* final consensus gene homeologous sequences……...……68

Table 2.8: DNA concentrations of *Pachycladon* species used for Illumina MiSeq and average library sizes obtained…….…71

Table 2.9: Percent identity values between homeologous sequences……...…72

Table 3.1: Molecular evolution tests for *Pachycladon* homeolog datasets……110

Table 3.2: Significant amino acid mutations and functionally important amino acids in the *Pachycladon* proteins GL1 and TTG1……114

Table 3.3: The top protein 3D model predicted by Phyre2 for GL1 homeologs, PDB top template information with GO annotations and *A. thaliana* GO molecular function annotation……...…….116

Table 4.1: *Pachycladon* species used in the study…….……148

Table 4.2: Selected reference genes for the study……..150

Table 4.3: qPCR primers used. 'Ref' indicates the reference genes primers……..…151

Table 4.4: qPCR program used in the LightCycler 480……...……..154

Table 4.5: qPCR primer efficiencies……...…159

Table 4.6: Unpaired 2-tail t-test results between all GOI (*TTG1, GL1, GL3* and *EGL3*) Type 1 and Type 2 homeologs of (a) *P. enysii* (*P.en*) seedling, (b) *P. cheesemanii* (*P. ch*) seedling, (c) *P. cheesemanii* (*P. ch*) mature leaf, (d) *P. fastigiatum* (*P.fs*) mature leaf and (e) *P. novae-zealandiae* (*P.nz*) mature leaf……..…160

Table 4.7: Unpaired 2-tail t-test results between (a) Type 1 homeologs and (b) Type 2 homeologs of all genes (*TTG1, GL1, GL3* and *EGL3*) between young *P. cheesemanii* (*P.ch*) and mature *P. cheesemanii*……...…..167

List of Appendices

Table A1.1: MIQE checklist for authors, reviewers, and editors……..…….46

A2.1 Statement of contribution……...……90

A3.1 Statement of contribution…….136

Figure A4.1: Standard curves for gene primer efficiencies……..……185

Figure A4.2: Melt curves of primer products……..……...……..187

Figure A4.3: Standard curves for cDNA dilution series of (a) young *P.enysii*, (b) mature *P.fastigiatum*, (c) mature *P.novae-zealandiae*, (d) young *P.cheesemanii* and (e) mature *P.cheesemanii*……...…….187

Table A4.1: *Pachycladon* cDNA dilution series used with TTG1-T1-1920F×1922R primer pair and their primer efficiencies…….……...……..188

Table A4.2: *Pachycladon* expression data (Cp values) for the genes of interest and reference genes……...……...……188

Figure A4.4: Distribution of Cp values of homeologous genes…….....…..193

Table A4.3: Gene expression analysis using qRAT……..…..…….193

Figure A4.5: qRAT Gene expression results of *TTG1*, *GL1*, *GL3* and *EGL3* homeologs in *P. enysii* (*P. en*) seedling, *P. cheesemanii* (*P. ch*) seedling, *P. ch* mature leaf, *P. fastigiatum* (*P. fs*) mature leaf and *P. novae-zealandiae* (*P. nz*) mature leaf……...……...…194

A4.1 Statement of contribution…….195

CHAPTER ONE

INTRODUCTION

This chapter describes the fundamentals of gene and genome duplication, discussing previous studies conducted on the fate of gene duplication in single genes, genomes, and genetic pathways. My research focuses on the trichome initiation genetic pathway in the allopolyploid study system *Pachycladon* (Brassicaceae) and consideration for the selection of this study system also is included in this chapter.

1.1 Gene duplication

A gene is referred to a DNA or RNA sequence that is capable of encoding functional molecules, either RNA or protein (Gerstein et al 2007). Duplication of a gene creates a redundant copy of the locus (Ohno 2013). The earliest observation of gene duplication dates back to 1936 when Bridges reported the doubling of a chromosomal band in a mutant of the fruit fly *Drosophila melanogaster* (Bridges 1936). The potential effect of gene duplication on evolution was suggested there. Later Ohno's seminal book in 1970, Evolution by Gene Duplication (Lee et al 2020, Ohno 1970), further popularized the idea of gene duplication. Gene and genome duplication studies were rapidly developed later with genome sequencing projects using improved high-throughput methods and population genetic modeling.

Gene duplication prevails in all three domains of life in bacteria, archaea, and eukarya, but is observed to be most common in plants in the eukaryotic category (Zhang 2003). It is thought to have been important in the evolution of these organisms, contributing to genome architecture, adaptation, and speciation (Panchy et al 2016). Gene duplication can provide new genetic material for mutation, drift (fluctuation of the frequency of alleles within a population over time), and natural selection to act upon, which is vital for the adaptation of new specialized gene functions in changing environments (Magadum et al 2013). Duplicated genomic regions and genome size differences between closely related species are some key evidence that shows gene and genome doubling have occurred during plant evolution (Adams

& Wendel 2005, Pellicer et al 2018). In fact, about 65% of annotated genes in all plant genomes have a duplicate copy (Panchy et al 2016).

The scale of gene duplication can be small or great including tandem and segmental duplications that often arise during DNA replication and recombination, and whole genome duplication or polyploidization. Mechanisms such as unequal crossing over and retroposition generally lead to segmental or tandem duplication (Flagel & Wendel 2009). Unequal crossing over could occur in chromosomal recombination when the DNA molecules from the two parents lie at non-identical locations. Hence, in unequal crossing over, a part of a gene, the entire gene, or several genes are linked with the chromosome, becoming a duplicated region. Therefore, unequal crossing over generates tandem duplicates, in which duplicated genes are linked in a chromosome. In retroposition, a mRNA is retrotranscribed into a cDNA and then inserted into the genome. In contrast to unequal crossing over, a duplicated gene generated by retroposition is usually unlinked to the original gene and the insertion position is more or less random (Zhang 2003). In contrast, chromosomal or whole genome duplications (polyploidy) are large-scale duplications that frequently occur in plants but infrequently in animals (Mable 2004, Román-Palacios et al 2021).

1.1.1 Polyploidy

Polyploidy is defined as having more than two complete sets of chromosomes per nucleus in a cell in an organism (Sattler et al 2016, Soltis et al 2003). The occurrence of polyploidy in plants was described more than 100 years ago by the work of Hugo de Vries based on autotetraploid mutants of *Oenothera biennis* and tetraploid *Digitalis mertonensis* (De Vries 1910, DeVries 1915, Vries 1901). Then Winkler generated the first artificial polyploid (Winkler 1916). Later, Kihara and Ono differentiated between allopolyploidy and autoplolyploidy and developed the two terms (Kihara & Ono 1926) (described later in 1.1.2). These early conceptions of polyploidy were described in the 1910s-1930s. Importantly, George Ledyard Stebbins, Jr. is often recognized as the pioneer of addressing the remarkable potential influence of polyploidy in plant evolution in the 1940s-1950s because he developed a model of

polyploid evolution with his coworkers that persisted for nearly half a century including many important concepts (Soltis et al 2014, Stebbins Jr 1950).

After the early descriptions of polyploidy, there was an era of genetic and cytological studies (1930s-1960s) using experimental polyploids in model systems. Pioneers such as Albert Blakeslee and Barbara McClintock explored the polyploids jimson weed and maize to study their chromosomal behavior and genetics (Blakeslee 1941, McClintock 1929, Satina & Blakeslee 1937). Such studies accompanied by many other studies identified gigas effects of neopolyploids such as increased cell size, slowed cell division and tissue development, increased organ size at maturity, and reduced fertility compared to diploid progenitors. Karyotype studies were conducted to observe the behavior of polyploids at meiosis. The experiments showed that many autopolyploids have random multivalent pairing (McCollum 1957, Morrison & Rajhathy 1960) whereas allopolyploids have bivalent pairing (Gerstel & Phillips 1957). Next in the 1950s-1970s, investigations of complex relationships among groups of hybridizing species and polyploid derivatives marked the era of biosystematic area of polyploidy research. Lastly, with the tremendous improvement of molecular tools and systematics, polyploidy research entered its molecular era (1980s-2000s) (Ramsey & Ramsey 2014, Soltis et al 2014).

With the availability of sequenced *A. thaliana* genome, Vision, and co-workers were able to find that *A. thaliana* has had at least three rounds of polyploidy (Vision et al 2000). From there, next-generation sequencing led to lots of new whole genome sequences for plants and eventually led to the conclusion that all the angiosperms have experienced at least one whole genome duplication (WGD) event in their ancestry (Farhat et al 2019, Jiao et al 2011, Van de Peer et al 2017), based on study systems such as wheat, maize, solanaceous crops tomato and potato, and so on (Blanc & Wolfe 2004b, Schlueter et al 2004).

1.1.2 Origins of polyploids

Traditionally, polyploids are classified into two types: autopolyploidy and allopolyploidy. Commonly, polyploidy arising within one or between populations belonging to a single species is called autopolyploidy, while polyploidy arising from different species or more highly diverged lineages is known as allopolyploidy. This is a little simplistic, however, and definitions that differentiate between those that form multivalents at meiosis (autopolyploids) and those that form bivalents at meiosis (allopolyploids) provide a more useful point of difference (Doyle & Sherman-Broyles 2017, Ramsey & Schemske 2002). Some examples of natural autopolyploids are sugarcane (*Saccharum* spp.) (Wang et al 2010), sweet potato (*Ipomoea batatas*) (Sumitomo et al 2019) and chives (*Allium schoenoprasum*) (Stebbins Jr 1950) and natural allopolyploids are salsifies (*Tragopogon* spp) (Ownbey & McCollum 1953), cotton (*Gossypium* spp) (Grant 1981) and *Brassica* (Soltis et al 2004, Song et al 1995).

There are several potential pathways to polyploid formation. The formation of unreduced reproductive cells (gametes) followed by mating is thought to be the key, but not the only pathway generating natural polyploids. Unreduced gametes can be fused with another unreduced gamete (bilateral polyploidization) or with a reduced gamete (unilateral polyploidization) (Sattler et al 2016). Bilateral polyploidization can produce both autopolyploids and allopolyploids by the union of two unreduced gametes from the same species or different species, respectively (Soltis et al 2003). Unilateral polyploidization generally produces an intermediary polyploid, a triploid which then backcrosses to a diploid (Sattler et al 2016). Figure 1.1 illustrates those possible pathways that allopolyploids and autopolyploids form. Somatic doubling is another route generating mixoploid chimeras or complete polyploids (Ramsey & Schemske 1998). This could take place in zygotic, embryonic, or sporophytic tissue when the cell cycle signaling is not in control, which is also identified as the most common process of polyploid formation (Tamayo-Ordóñez et al 2016).

Figure 1.1: Possible pathways of polyploid formation. For simplicity, not all possible paths are shown here. Triploids, for example, are shown contributing to autotetraploids but they can also contribute to allopolyploids. For each ploidy form, the haploid genome is represented by a coloured circle or oval inside the beige-filled nuclear shape. Genomes illustrated by ovals reflect the increased gene number resulting from the retention and subfunctionalization of duplicates during diploidization. Circles or ovals of different colours represent diverged genomes. Highly unstable ploidy forms have dashed nuclear contours. A and B represent genome types and N is the gametic chromosome number. The figure was extracted from Comai (2005).

1.1.3 Polyploidy and plant evolution – Molecular consequences of polyploidy

Post-polyploidization brings rapid changes in genome structure, gene content, gene expression, methylation, morphological and physiological adaptations, and speciation (Moghe & Shiu 2014, Soltis et al 2014).

1.1.3.1 Chromosomal loss, fusion and rearrangements

Some polyploids undergo rapid genome rearrangements after the polyploidization, including chromosomal inversions, translocations, duplications, insertions, and deletions (Chen & Ni 2006, Hufton & Panopoulou 2009). Generally, chromosomal rearrangements are detected

in a genome by changes in the order, size, or position of various genomic markers. In paleopolyploids, the level of genomic rearrangements can be analyzed by studying the level of conserved synteny (gene order) in the segments identified between related species, if fully sequenced genomes are available (Paterson et al 2012). In contrast, in recent polyploids, cytogenetic methods and changes in molecular markers are used to infer the occurrence of genomic rearrangements (Hufton & Panopoulou 2009).

Various mechanisms were suggested to stimulate genomic rearrangements post polyploidization, such as transposable element activation, meiotic/mitotic defects, and DNA deletion (Comai 2000, Comai 2005, Tate et al 2005). Some examples of chromosomal rearrangements include aneuploidy (alterations in the number of chromosomes), as a result of meiotic dysfunction (studied in *Arabidopsis* and *Asteraceae*) (Lim et al 2008, Madlung et al 2005), homeologous recombination (as observed in *Brassica* allopolyploids) (Gaeta et al 2007, Pires et al 2004, Udall et al 2005), and genomic deletions (reported in natural and synthetic wheat polyploids) (Levy & Feldman 2004).

1.1.3.2 Duplicate gene loss and retention

Even though polyploidization is a recurrent process that has taken place in all plant lineages (Dodsworth et al 2016, Soltis & Soltis 1999), generally the present gene numbers and genome sizes in most lineages are far less than what would be expected after multiple rounds of genome doubling, which means a great deal of gene loss occurs after polyploidization events (Adams & Wendel 2005, Wang et al 2021a). One of the entire homeologous copies (duplicated copy) or regulatory site of a homeologous copy from a duplicated gene pair can be lost, referred to as fractionation. Further, these losses may be random concerning parental genome, or directional - a process referred to as biased fractionation (Soltis et al 2015). For example, the *B. rapa* genome has demonstrated fewer gene losses in one of its three subgenomes compared with the others; copies from that subgenome are also expressed higher than in the others (Tang et al 2012, Wang et al 2011). There are a lot of mechanisms suggested for gene deletions such as intra-chromatid recombination followed by small RNA-mediated epigenetic repression,

mutational decay, pseudogene pathway, and transposon activity (Cheng et al 2018, Freeling 2009, Langham et al 2004, Wendel et al 2018).

While some genes tend to be lost after genome doubling, other genes are preferentially retained in the genome. Several studies have revealed that the differential retention of homeologs is a process that is subject to natural selection and not random (Adams & Wendel 2005) and particular patterns at sequences, expression, and molecular function impact gene retention (Ganko et al 2007, Jiang et al 2013). In *Arabidopsis*, some classes of genes, such as those involved in transcription and signal transduction, have been preferentially retained in duplicate, whereas other classes, including those involved in DNA repair and organellar proteins, have been preferentially reduced to single copy (Blanc & Wolfe 2004a, Blanc & Wolfe 2004b). Further, overall, genes with more pleiotropic effects have been found to be preferentially retained compared to genes with low pleiotropic activity, suggesting retention of highly pleiotropic genes may be advantageous in genetic redundancy (Láruson et al 2020, Zhang 2012) and maintaining dosage sensitivity (Ascencio et al 2021, Cui et al 2006). But in some cases, retention of only one pleiotropic copy could be more beneficial because two copies cause dosage imbalance and metabolic burden demanding more resources and energy (Ascencio et al 2021, Qian et al 2012).

In the context of genetic networks, it has been suggested that other interacting genes influence the retention of duplicated genes because many other genes directly or indirectly influence the function of any one duplicated gene (Panchy et al 2016).

1.1.3.3 Functional divergence of retained duplicated genes

When both copies of a duplicated gene are retained, one duplicated gene copy could retain the original function and the other one can acquire a new function (neo-functionalization), or each copy may retain some part of the original functions of the gene (sub-functionalization), or sometimes, both copies can retain their original function (Adams & Wendel 2005).

To go through the process of sub-functionalization, it is highlighted that the ancestral gene must have at least two important functions to perform with multiple expression domains (Flagel & Wendel 2009, Zhang 2003). Cheng et al (2018) further elaborate sub-functionalization as a two-mutation process that comprises one loss-of-function mutation in each duplicate but in different cis-acting subfunctions, thus both duplicates are required to be retained to keep the combined ancestral function. Li and Noll (1994) suggested an example for this showing triplicated *Drososphila* genes having conserved protein but distinct developmental functions.

Neo-functionalization is the origin of a new beneficial function in one copy of the duplicated gene while the ancestral function is retained in the other copy (Adams & Wendel 2005). Neo-functionalization has been identified as a process undergoing positive selection whereas sub-functionalization undergoes purifying selection (Roulin et al 2013). Therefore, neo-functionalization selects beneficial mutations, which makes it a rather slow and rare evolutionary mechanism compared to sub-functionalization (Freeling 2008, Freeling 2009). This may explain why recent polyploids have revealed only a few cases of new functions in duplicate gene copies while ancient paleopolyploids show more (Blanc & Wolfe 2004a). For instance, (Roulin et al 2013) study reports only a few cases of gene copies of neo-functionalization (around 4%) compared to approximately 50% sub-functionalized paralogs in Soybean (*Glycine max* L.) genome.

Meanwhile, one of the duplicated gene copies could turn out to be non-functional (pseudogenization) after mutational inactivation, such as premature stops or frameshift mutations. These pseudogenes are sometimes called “defunct” genes or “junk” DNA (Wang et al 2012, Zou et al 2009).

1.1.3.4 Differential gene expression

In allopolyploids, the combination of two diverged genomes is expected to create a transcriptome shock, which leads to rapid changes in gene expression patterns (Hegarty et al 2006). Homeologs may be expressed at equal levels, or unequal levels, or one copy may be

silenced completely (Adams & Wendel 2005). For example, *A. thaliana* duplicates derived from more recent α and older β polyploidization events were found 57% and 75% diverged in expression respectively (Blanc & Wolfe 2004a).

Silencing that arises at the onset of polyploidy may be epigenetically induced (changes in cytosine methylation and histone modification), due to chromatin structure changes, or due to small RNAs and RNA interference (RNAi) (Liu & Wendel 2003). Further, the aggregation of point mutations can lead to gene silencing in later polyploid generations (Adams & Wendel 2005).

The selective reasons behind gene silencing can be to maintain appropriate gene dosage balance and to maintain the requirements imposed by regulatory networks (Edger & Pires 2009). When it comes to regulatory networks, for example, more maternal gene copies involved might be expressed while more paternal gene copies are silenced (homeologous silencing), if the maternal copies have better interaction with relevant proteins and enzyme-substrate interactions due to having different amino acid sequences than the paternal copy. They can be also subject to natural selection and later evolutionarily stabilized in natural polyploids. Alternatively, the silencing of some genes can be functionally neutral if it happens as a side effect of other mechanistic processes that take place in the cell (Adams & Wendel 2004, Adams & Wendel 2005, Soltis et al 2014). Furthermore, the literature reports polyploid gene silencing as a random process for some genes and seemingly deterministic for others. For example, parallel silencing of the same duplicate gene copy in multiple polyploid genotypes or lines has been observed in some systems (Adams et al 2004, Buggs et al 2012, Wang et al 2004a), but more random patterns of gene silencing across genotypes, lines, and generations have been reported in others (Joly et al 2004, Pires et al 2004).

Generally, two phenomena have been described to explain parental genome dominance in the duplicated transcriptome, which are termed homoeolog expression bias (Chen 2007) and expression level dominance (Rapp et al 2009). Homeolog bias refers to the preferential expression of one homoeolog relative to the other, which has been observed in many allopolyploids, including *Arabidopsis* (Chang et al 2010), *Brassica* (Gaeta et al 2007),

Tragopogon (Buggs et al 2010) and others. Expression dominance describes the phenomenon where the total expression of homeologs for a given gene in an allopolyploid is statistically equivalent to the expression level of that gene in only one of the parents. This also has been described in many studies including cotton (Rapp et al 2009), coffee (Bardil et al 2011) and wheat (Qi et al 2012). Preferentially expressed maternal genes are involved in interacting with plastid genes to regulate cytonuclear interactions to encode several plastid macromolecular protein complexes (Gong et al 2012). However, these fine-tuned interactions can be disrupted when the allopolyploid genome is hybridized with two divergent species and doubled with gene numbers. Therefore, maternal expression bias on the genes involved in plastid protein complexes has been hypothesized and shown in previous experiments. For example, Gong et al. found a consistent transcriptional bias of the maternal Rubisco gene homeologs (encoding a cytonuclear enzyme complex) in four natural allopolyploids, *Arabidopsis*, *Arachis*, *Brassica*, and *Nicotiana*.(Gong et al 2014, Grover et al 2012). Differential evolution of Rubisco maternal and paternal homeologs was also reported between homoeologous genomes of *Arabidopsis suecica*, *Arachis hypogaea*, and *Nicotiana tabacum* allopolyploids (Gong et al 2014, Gong et al 2012). However, another study demonstrated that cytonuclear interactions in *Brassica* were found to be stable showing higher retention of the genes from both parents involved in cytonuclear interactions without signs of any biased expression (Ferreira de Carvalho et al 2019).

1.1.3.5 Diploidization

Although previous work has shown that all angiosperms have undergone one or more episodes of polyploidization during evolution (Blanc & Wolfe 2004b, Jiao et al 2011, Van de Peer et al 2017), there is also widespread evidence for subsequent diploidization due to genomic rearrangements, gene silencing, and genomic downsizing (Soltis et al 2003). Modern molecular approaches such as chromosome painting (GISH, FISH), genetic mapping, and comparative genomics have provided evidence for genomic rearrangements following polyploidy. For instance, despite the fact that *Arabidopsis thaliana* and *Zea mays* are

paleopolyploids, currently, their meiotic behavior is that of functional diploids (Gaut et al 2000, Vision et al 2000).

Primarily, illegitimate recombination (non-homologous recombination) and unequal intra-strand homologous recombination are considered to be mechanisms for genome downsizing (Soltis et al 2015). Chromosomal segregation aberrations happen through these mechanisms leading to extensive aneuploidy. However, compensating aneuploidy has been described as a mechanism aiming for dosage balance. For example, in *Brassica napus*, despite the high rate of aneuploidy, the number of copies of a particular chromosome is frequently maintained; e.g., the loss of one copy of a parental chromosome 1 is usually associated with the gain of the same chromosome from the other genome, resulting in a 1:3 ratio instead of the initial 2:2 ratio of homeologs (Xiong et al 2011).

1.1.4 Ancestral polyploidization events related to the Brassicaceae lineage

Two ancient gene duplication events have been discussed in the literature to explain the radiations of seed plants and angiosperms, approximately 319 and 192 million years ago, respectively (Jiao et al 2011). Following these events, most flowering plant lineages experienced one or more further ancient polyploidization events (Cui et al 2006, Schranz & Mitchell-Olds 2006).

Markedly, genome evolution and divergence studies were carried out using the model mustard *Arabidopsis* because of its three ancient polyploidization events (Bowers et al 2003) as mentioned earlier. It is accepted that the ancient *A. thaliana* α whole genome duplication event pre-dated its divergence from *Brassica* around 14.5–20.4 million years ago (Myr), but post-dated its divergence from *Malvaceae* around 83-86 Myr, leading to the radiation of the mustard (Brassicaceae) family (Bowers et al 2003, Jiao et al 2011, Mandáková et al 2017, Van de Peer et al 2009a). A second widely known paleopolyploidy event is the core Brassicales-specific *A. thaliana* β whole genome duplication which was suggested to occur around 170–235 Myr ago (Bowers et al 2003). A third paleopolyploidy event γ (inferred as a hexaploidy event resulted

from two WGD events that occurred in close succession (Jiao et al 2012), was interpreted to have taken place before eudicot and monocot divergence (divergence shared by core-eudicots and *Oryza sativa*) around 150 Myr (Bowers et al 2003, Simillion et al 2002, Van de Peer et al 2009a).

1.1.5 Homeolog fate in a genetic pathway context

Although much research has been done on homeolog fate either in a whole genome context or focusing on single genes following WGD (Flagel & Wendel 2009, Liu & Zhang 2019, Roulin et al 2013), there is not much work on homeolog fate within a system such as in the context of a genetic pathway. Considering homeolog function in a system such as a genetic pathway will help understand how duplicate gene copies collectively influence molecular functions and related traits in plants (Panchy et al 2016). Therefore, the present study focuses on homeolog fate in the Brassicaceae trichome cell fate pathway.

In a genetic pathway, a series of genes hierarchically influence each other's function (Shimeld 1999). Following whole genome duplication, all of the components in such a genetic pathway are doubled. The gene dosage balance theory states that duplicate gene copies are preferentially co-retained without silencing one of the copies when their products participate in protein complexes and so the stoichiometry among interacting gene products will be maintained (Conant et al 2014). Many previous studies have shown that gene dosage balance acts on multiprotein complexes and signal transduction networks (Aury et al 2006, D'Hont et al 2012).

But this is not always the case and some of the pathway components may be retained in duplicate while others end up reduced to a single copy. A study from invertebrates (*Caenorhabditis elegans* and *Drosophila*) and jawed vertebrates, on components of the wingless and hedgehog signaling pathways has shown variations in gene retention of the pathway components (Shimeld 1999). This study also states that the genetic components of a pathway could acquire different functions after gene duplication, based on its structure, function, and network position. Moreover, Kaltenegger et al (2018) report a highly

heterogeneous pattern of gene retention after WGD in core genes of the cytokinin signaling pathway. Further, they have also observed the downstream elements in the pathway were likely to be retained while upstream elements seem to be lost, following the agreement that upstream genes in a biochemical pathway evolve more slowly than downstream genes (Lu & Rausher 2003, Rausher et al 1999).

Duplicate gene copies in a genetic pathway may be impacted by homoeologous biased fractionation or silencing of one whole parental sub-genome, resulting in genes expressed from a single parent as discussed in the whole-genome context above. However, maintenance of duplicated gene copies in a genetic pathway or network is not an area adequately understood yet. The present study tries to address this question further, using a well-annotated genetic pathway by analyzing the homeologous components of the pathway through their sequences and gene expression.

1.2 Brassicaceae epidermal cell fate pathway

The plant epidermal cell fate pathway has been largely resolved through decades of work in the model plant *Arabidopsis thaliana*. As such, this network of genes that determine epidermal cell fate makes for a good candidate system for studying the outcome of whole genome duplication events at the level of the genetic pathway.

The regulatory pathways for epidermal cell differentiation consist of a core transcriptional network with common structures in a transactivation complex including multiple MYBs, basic helix-loop-helix (bHLH) transcription factors, and a single WD40 class protein (TTG1). A protein complex is the combination of proteins which interact with each other (Wang et al 2021c). Common downstream targets of this pathway are the homeobox gene *GLABRA2* (*GL2*) and WRKY transcription factor gene *TTG2* (Chu & Corey 2012, Pesch et al 2014). Hence, *GL2* and TTG2 genes in the pathway are targeted by the activation complex and are negatively regulated by a family of R3 MYB proteins, that competitively inhibit the R2R3 proteins from binding with the bHLH proteins in the complex (Doroshkov et al 2019).

In the root epidermis, the activation complex promotes non-hair cell differentiation, whereas in the leaf epidermis, it promotes both cell pattern formation and trichome differentiation, while producing long hypocotyl cells without stomata in hypocotyl epidermis, mucilage in the seed coat and anthocyanin biosynthesis in various plant parts. The components, whether common or specified, which are involved in each of these epidermal cell fate pathways are shown in Figure 1.2 below. As presented in the figure, mainly TTG1 and bHLH transcription factors EGL3 and GL3 are active in multiple epidermal cell fate pathways and therefore have pleiotropic effects.

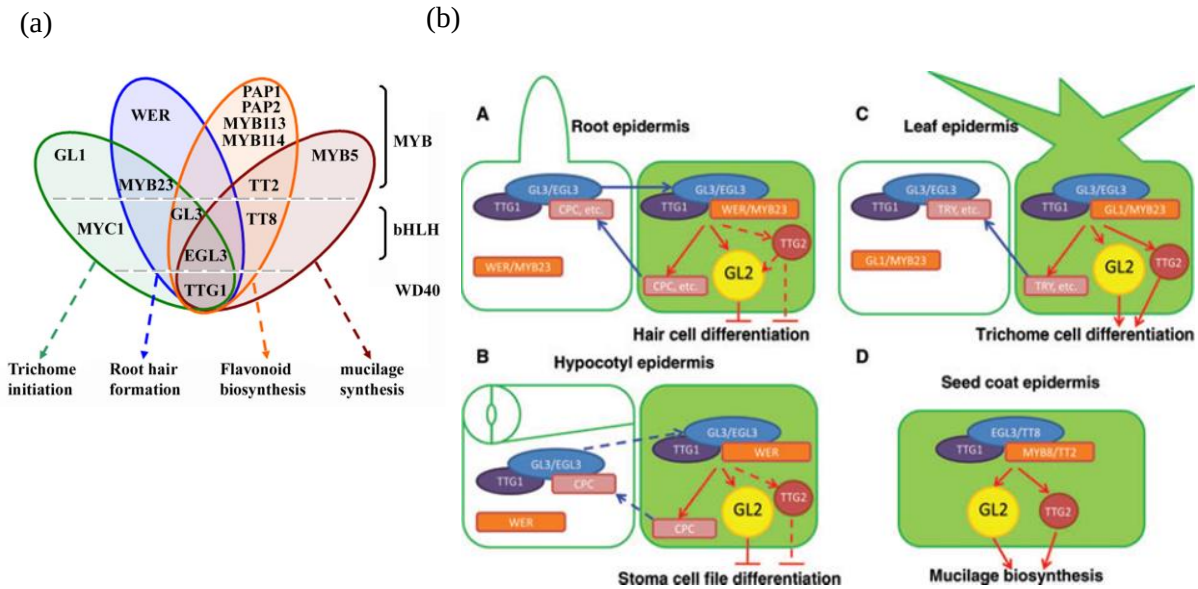


Figure 1.2: Proteins involved in forming the common activation complex in different epidermal cell fate determination pathways in *Arabidopsis*. **(a)** Transcription factors and proteins belong to R2R3 MYBs, bHLHs, and WD40 groups involved in the common activation complex in different epidermal cell fate determination pathways in *Arabidopsis* reproduced from Chen and Wang (2019). Some of the components are common in all or multiple activation complexes and some are present in specific activation complexes only (Although this extracted figure shows that MYB5 is specifically involved in mucilage synthesis, it's secondary role is also described in leaf and stem trichome morphogenesis (Gonzalez et al 2009, Li et al 2009b) **(b)** Regulatory pathways for epidermal cell differentiation reproduced from Chu and Corey (2012) Schematic illustrations of pathways upstream of *GL2* and *TTG2* for root epidermis (A), hypocotyl epidermis (B), leaf epidermis (C) and seed coat epidermis (D). Red arrows and T-bars indicate positive and negative transcriptional regulation. Blue arrows indicate protein movement between cells. Dashed-line arrows and T-bars indicate assumed regulation and movement. Cells that strongly express *GL2* are colored in green.

The R2R3-type MYB transcription factor constitutes one of the largest gene families in angiosperms, where the R2R3 domain is responsible for DNA binding, promoter specificity, and interaction with other co-factors (e.g. bHLHs). Chen and Wang (2019) state that nine R2R3-type MYBs actively participate in epidermal cell differentiation in *Arabidopsis*: GL1, MYB23, WER, PAP1, PAP2, MYB113, MYB114, TT2 and MYB5. Related to the R2R3 MYBs, *Arabidopsis* has eight R3-type MYBs active in the epidermal cell fate pathway: TRY, CPC, ETC1, ETC2, ETC3/CPL3, TCL1, TCL2 and MYBL2. They are primarily involved in binding with bHLHs (Zimmermann et al 2004) and cell-to-cell movement (Kurata et al 2005, Wang & Chen 2014), but they lack an activation domain, which is typically present in transcription factors (Wang & Chen 2014).

bHLH transcription factor superfamily is also one of the largest transcription factor families in *Arabidopsis* (Toledo-Ortiz et al 2003, Zhao et al 2012a). Four bHLHs in *Arabidopsis* participate in epidermal cell fate pathways (GL3, EGL3, MYC1, and TT8) (Chen et al 2019) and they have a conserved DNA binding domain (Chezem & Clay 2016). bHLH is the key protein acting as the linker between the WD40 protein TTG1 and R2R3-MYB proteins in making the transactivation complex because no direct interaction has been identified between TTG1 and R2R3-MYB proteins (Tominaga-Wada et al 2011). bHLHs possess a C-terminal DNA binding domain important in binding with the downstream gene *GL2*, and the N-terminus for binding with TTG1 and R2R3-MYBs (Zhao et al 2008).

The WD40 protein TTG1 contains four WD repeat motifs that facilitate protein-protein interactions, but with no recognizable DNA binding motif, transcriptional activation domain or nuclear localization signal (Walker et al 1999, Zhao et al 2008). But by interactions with bHLHs, TTG1 can act as both a protein interaction factor in the cytoplasm and a transcriptional co-regulator in the nucleus (Zhao et al 2008).

1.2.1 Cell differentiation in the leaf epidermis

1.2.1.1 Plant trichomes and the origin of *Arabidopsis* trichomes

Plant trichomes are specialized epidermal outgrowths that play many roles in plant defense. They protect plants from adverse environmental conditions and other external threats, such as herbivorous insects and toxic chemicals, and generate and store useful compounds important for plant development which also can be utilized for human use and disease treatment (Wang et al 2019). Some examples of human uses of trichomes include as raw material for natural fiber (trichomes in the seed coat of cotton) (Rinehart et al 1996) synthesis of polysaccharides, proteins, polyphenols, and terpenoids from *Humulus lupulus* and *Artemisia annua* trichomes which can be used for producing medicine, food additives, herbicides, and resin (Markus Lange & Turner 2013, Singh et al 2016) and production of artemisinin, an antimalarial drug using the trichomes of *A. annua* (Singh et al 2016).

Morphologically, as large epidermal cells that are easily accessible, trichomes have become an ideal model system for genetic and cell biological studies. The *Arabidopsis thaliana* trichome is a large single non-secreting epidermal outgrowth cell, composed of one to three or four branches that initiate very early in leaf primordium formation from protodermal cells at the shoot apical meristem (Morohashi & Grotewold 2009). The epidermal cells committed to a trichome fate do not proceed to cell division while surrounding epidermal cells continue to divide. Instead, they proceed with an average of four DNA replication cycles which is referred to as endoreduplication, which facilitates expansion away from the leaf surface and successive branching events (Grebe 2012, Hülskamp et al 1994, Schwab et al 2000). Based on cellular interactions that occur between neighboring epidermal pavement cells, trichomes are regularly distributed on the leaf surface; which individual cells become trichomes in this pattern is decided by a combination of endogenous developmental controls and external signals (Balkunde et al 2010, Hauser 2014, Ishida et al 2008, Okamoto et al 2020).

1.2.1.2 Trichome initiation pathway in *Arabidopsis*

The regulation of trichome initiation involves the R2R3-type MYB transcription factor GLABRA1 (GL1), MYB23 and MYB5, the bHLH transcription factors GLABRA3 (GL3), ENHANCER OF GLABRA3 (EGL3) and MYC1, and the WD40 protein TRANSPARENT TESTA GLABRA1 (TTG1), which bind to form transactivation complexes (Figure 1.2 (b) C) (Chu & Corey 2012). Among the primary targets of this transactivation complex are *GL2*, *TTG2*, and *SIM*, which are required for endoreplication (Grebe 2012). The bHLH, GL3, of the transactivation complex, can directly bind with the promoters of R3-type MYB genes *TRIPTYCHON* (*TRY*), *CAPRICE* (*CPC*), and *ENHANCER OF TRY AND CPC 1* (*ETC1*) in trichome cells and then their proteins are moved into neighboring cells, and thus suppress the transactivation complex by competing with GL1 for binding with the bHLHs. As a result, *GL2* expression is inhibited and trichome initiation does not occur (Chu & Corey 2012, Grebe 2012) in the neighboring cells. This mutual interaction between positive and negative regulators is thought to explain the regular trichome patterning observed (Balkunde et al 2010).

Many studies have investigated how these transcription factors and proteins physically interact with each other to form the transactivation complex, and it has helped to understand how each component's expression affects others and, in turn, the whole mechanism. The previous experiments by (Payne et al 2000, Zhang et al 2003) found that GL3 has three distinct protein-protein binding domains: one to bind with MYB proteins, another for TTG1, and a third to interact with itself (homodimerize), but TTG1 and GL1 were reported to not directly interact (Payne et al 2000, Zhao et al 2008). EGL3 was found to have the same protein-protein interactions as GL3 (Payne et al 2000, Zhang et al 2003).

Researchers have revealed that loss of expression of *GL1* and *TTG1* inhibit trichome differentiation while *gl2* mutants form aborted trichomes with aberrant cell expansion, indicating that *GL2* is a positive regulator for part of trichome differentiation while *GL1* and *TTG1* are crucial positive regulators for trichome initiation (Chu & Corey 2012, Larkin et al 1994, Larkin et al 1999, Wang et al 2019). *TTG2*, the other important downstream gene, has been shown to directly regulate *TRY* by forming a protein complex with TTG1, and GL3. Thus,

TTG2 also enhances the activity of TTG1 and GL3, by physically binding with TTG1 and then associating with GL3, via the interaction with TTG1 (Pesch et al 2014).

It is reported that the transcription factor GL3 is mainly responsible for endoreduplication in trichome cells and thus, affecting trichome cell size and branching (Payne et al 2000, Wang et al 2021b). EGL3 has an overlapping role with GL3, and *gl3* and *egl3* double mutant plants are totally glabrous (Zhang et al 2003). When both genes *GL3* and *EGL3* are overexpressed in a *ttg1* mutant background, the trichome defect caused by *ttg1* is suppressed as well (Zhang et al 2003). *AtMYC1* in *Arabidopsis* was identified as a transcriptional regulator influencing trichome density through its natural allelic variation (Symonds et al 2011).

Other than these key regulatory factors, some upstream transcription factors, phytohormones, and epigenetic modifications affect trichome development in plants. For example, mainly the phytohormone gibberellic acids (GAs) and GIS transcription factors are involved in the GA pathway and have shown some roles in trichome development (Payne et al 2000, Wang et al 2019, Zhao et al 2008).

With any whole genome duplication (WGD) event, all genes that make up a particular genetic pathway will be duplicated; however, the genetic pathway context is rarely examined in studies considering the effects of WGD on gene duplicate fate. The key genes involved in the trichome initiation transactivation complex *TTG1*, *GL1*, *GL3*, and *EGL3* will have been duplicated in each whole genome duplication they have faced, and this thesis work focuses on their fates in the allotetraploid *Pachycladon* genus (*Brassicaceae*).

1.3 *Pachycladon* (New Zealand rockcress) as a study system

Pachycladon (Brassicaceae) is a genus of perennial herbs that are native to New Zealand and Australia (Tasmania), where they grow mostly on rocky substrates. The somatic chromosome number of *Pachycladon* species is 2n=20. Based on (Joly et al 2009), the genus has a recent allopolyploid origin. The lineage was formed via hybridization of two highly divergent species. At this point, it is unclear whether this would have involved direct formation

through via unreduced gametes or through diploid hybridization, followed by allopolyploidization. Following polyploidy, the lineage underwent rapid radiation and speciation, evident by its wide range of morphological diversity and habitat preference among species (Figure 1.3) (Mandáková et al 2010). It has emerged as a model plant group for the study of polyploidy, in part due to the highly diverged nature of the progenitors, but also due to the relatively close relationship with the model plant *Arabidopsis thaliana* (Yogeeswaran et al 2011). This close relationship between the *Pachycladon* and *Arabidopsis* genera is underlined by their placement in the polyphyletic tribe *Camelineae* (Al-Shehbaz et al 2006, German et al 2009), and specifically through the formation of an artificially produced intergeneric hybrid between *A. thaliana* and *P.cheesemanii* (Heenan et al 2008).

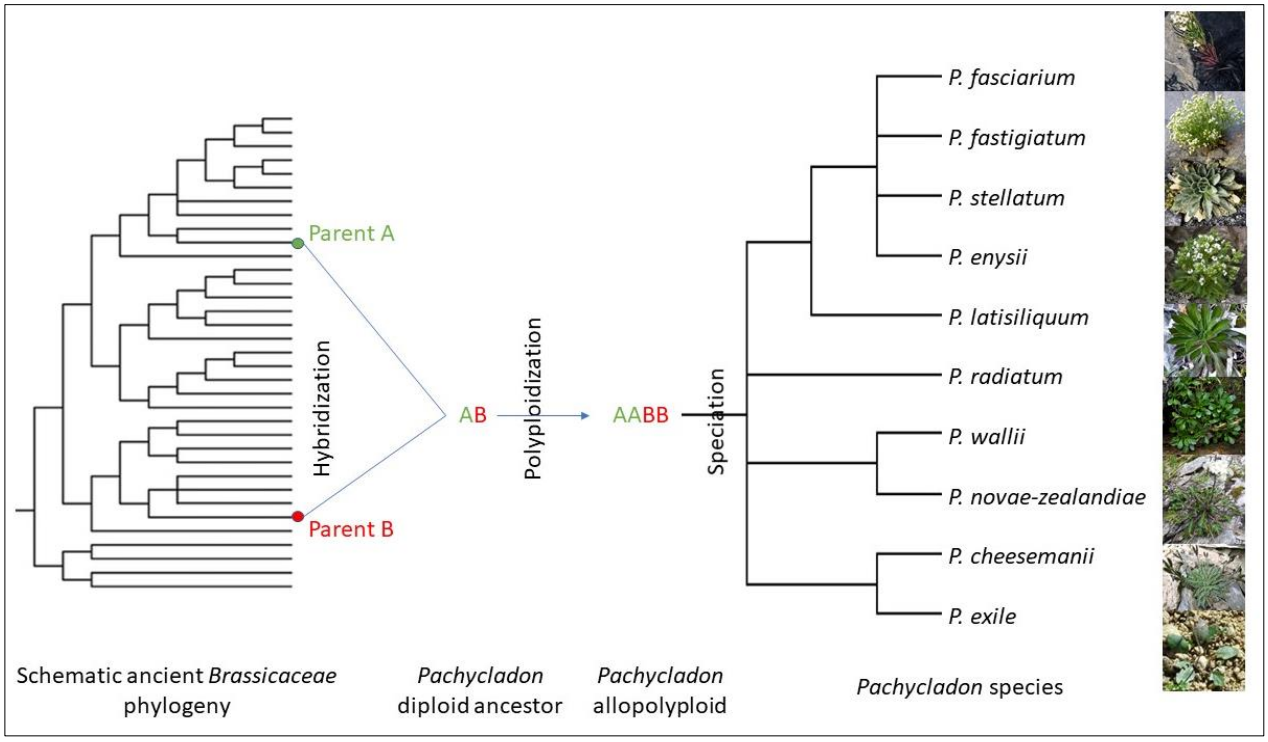


Figure 1.3: A schematic diagram showing ancient origin of *Pachycladon*, hybridization, allopolyploidization and speciation. *Pachycladon* species tree was extracted from Mandáková et al (2010) and modified.

1.3.1 Basic botany and distribution of *Pachycladon*

There are 10 recognized species of *Pachycladon*, consisting of nine species endemic to New Zealand and one species endemic to Tasmania (*Pachycladon radicatum*). All New

Figure 1.4: (A) Distribution of all New Zealand *Pachycladon* species **(B)** NeighborNet network of inter-organism genetic distances computed from 10 single-copy nuclear genes with the POFAD algorithm; figure reproduced from Joly et al (2013).

Zealand species are distributed on the South Island and cluster into three primary groups based on their morphology and habitat preferences (Heenan & Mitchell 2003). Phylogenetic analysis of morphological characters reassured this grouping and included a fourth group comprised solely of *P.radicatum*, which is the Tasmanian species (Heenan & Mitchell 2003, Yogeeswaran et al 2011). Figure 1.4 below shows the distribution of New Zealand *Pachycladon* species on the South Island of New Zealand and genetic relationships among the New Zealand species. *Pachycladon cheesemanii*, which is widely distributed in the eastern part of the South Island (Heenan & Mitchell 2003), is the most abundant species in the group.

Pachycladon species are rosette-forming herbaceous plants with a perennial life history and glabrous lamina or simple/branched trichomes, terminal or lateral inflorescences, and narrow siliques (Yogeeswaran et al 2011). All species have bisexual flowers comprised of four white petals, six stamens, and one pistil. These plants flower from early to mid-summer generally between September and January. The morphology and growth habits of the four species under study in this thesis are shown in Figure 1.5. Although we do not measure any of

these phenotypes in *Pachycladon* plants in this study, we record the differences in trichomes where appropriate to explain our results.

Some species are able to self-fertilize such as *P.cheesemanii*, *P.exile*, and *P.novae-zealandiae* with selfing floral morphology and low pollen:ovule ratios, while some are encouraged to outcross such as *P.fastigiatum*, *P.stellatum*, and *P.latisiliquum* with associated changes to floral morphology and higher pollen:ovule ratios (Garnock-Jones 1991, Heenan 1999, Mitchell & Heenan 2002, Yogeeswaran et al 2011). Further, interspecific hybrid combinations have been reported in the wild, and F1 hybrids can be generated artificially (Bicknell et al 2009, Heenan 1999, Yogeeswaran et al 2011) as might be expected for a recently radiated group of species (Yogeeswaran et al 2011).

Figure 1.5: Morphology and growth habit of (a) *P. cheesemanii*, (b) *P.enysii*, (c) *P.fastigiatum* and (d) *P. novae-zealandiae*, extracted from Yogeeswaran et al (2011).

To cultivate *Pachycladon* species, typically, fresh seeds rapidly germinate in damp conditions, while 2–3-week-old seeds need to go through stratification at 4 °C for 7 days and then germination usually takes place around 7-12 days after stratification. Also, propagation can be made artificially by cuttings, which prefer greenhouse conditions to grow (temperatures below 20-22 °C, good airflow and overhead shading preferred ~350 µmols^{-1} m^{-2}) (Yogeeswaran et al 2011).

1.3.2 Allopolyploid origin of *Pachycladon* within the *Brassicaceae* phylogeny

The allopolyploid origin of *Pachycladon* is said to be one of the most extreme examples of successful hybridization between distantly related parents according to (Joly et al 2009). McBreen and Heenan (2006) first cited the allopolyploid origin of *Pachycladon*, exploring the duplication at the glyceraldehyde 3-phosphate dehydrogenase (*gapC*) gene, which

revealed greater sequence divergence between homeologs within species than between the same homeolog among species (McBreen & Heenan 2006, Yogeeswaran et al 2011).

Later, the study of Joly et al (2009), investigated five single-copy nuclear gene markers (*CHS*, *PRK*, *MS*, *CAD5*, and *MtN21*), reporting two distantly related copies in each *Pachycladon* species. Here, the *CHS* (Chalcone synthase) phylogeny for *Brassicaceae* indicated that one copy of the gene that belongs to *Pachycladon* (referred to as genome A from here) is derived from the Arabidopsoid lineage, closely related to *Crucihimalaya*, *Transberingia*, and *Boechera*, while the other gene copy (referred to as genome B from here) is located at a more distant position within the *Brassicaceae* phylogeny. The B copy was identified as close to the split of the *Arabidopsis* and *Brassica* lineages, but closely related species are not known because of the uncertainty of the position. A later report (Zhao et al 2010) inferred the Brassicaceae phylogeny using *CHS* sequences and reported one *Pachycladon* copy nested in lineage I and the other copy nested within the *Smelowskiea* clade, which is also located in lineage I, but at a deeper position; however, the affiliation with the *Smelowskia* clade had very weak support. In the Brassicaceae phylogeny inferred by Huang et al (2015), *Pachycladon* clades were found to be sister to either *Boechera* or *Smelowskia*, analogous to the Zhao et al (2010) study.

Researchers have utilized the *Pachycladon* chloroplast genome to investigate its maternal inheritance. In one of the attempts, the chloroplast gene *rbcL* was used to compare *Pachycladon* among 52 other Brassicaceae representative species and found that *Pachycladon* was placed at the base of the *Arabidopsis* lineage, which is relatively distantly positioned from the genera *Transberingia* and *Crucihimalaya*, in a very similar position to that of B genome copy obtained with the nuclear genes' phylogeny, that mentioned above (Joly et al 2009). More recent studies (Hendriks et al 2023) presenting the most complete Brassicaceae phylogeny to date, with 60 plastome genomes and 265 genera, also places the *Pachycladon* maternal contribution outside of the clade that houses *Arabidopsis*, *Camelina*, *Capsella*, *Turritis*, and *Erysimum* groups, similar to the Joly et al (2009) study. However, another study involving four

chloroplast genes (*rbcL*, *nad4*, *mat*K, *ndh*F) by Mandakova et al (2010), which was analyzed with 55 representative *Brassicaceae* species placed *Pachycladon* within a clade with the other tribe Camelineae species such as *Boechera*, *Crucihimalaya*, and *Transberingia*, interpreting the *Pachycladon* genome A progenitor as the likely maternal parent of the genus. Overall, there is more recent comprehensive data to suggest *Pachycladon*'s maternal parent is more similar to Joly et al (2009)'s B genome copy.

The Mandakova et al (2010) also implied that the two *Pachycladon* nuclear copies are not as highly divergent as Joly et al (2009) states, with a finding that both copies are in the Camelineae lineage. According to the phylogenetic tree and analysis from Joly et al (2009), the two nuclear genomes diverged ~8.18 Mya, and the hybridization event that formed *Pachycladon* occurred 0.8-1.61 Mya during the Pleistocene. Therefore, further studies are required to elucidate the origin of *Pachycladon* genome copies.

It is clear that *Pachycladon* underwent an adaptive radiation due to niche shift following the hybridization event resulting in Tasmanian *P.radicatum* and New Zealand *Pachycladon* species (Joly et al 2014, Yogeeswaran et al 2011). Notably, *Pachycladon* is suggested as a remarkable example of adaptive radiation due to its high estimated diversification rate, distinctive ecological niches of species, and evidence for ecological speciation such as substrate specificity (Joly et al 2014). It is also suggested that this adaptive radiation was specifically triggered by the occurrence of hybridization and allopolyploidy (Joly et al 2009).

Further, *Pachycladon* is thought to have undergone recurrently occurred three *Brassicaceae* paleopolyploidy events identified (described in 1.1.4), followed by an additional mesopolyploidy event (Mandáková et al 2017). At least 13 independent mesopolyploidy events across *Brassicaceae* were discussed, and one of these events was identified at the divergence of two genera *Pachycladon* and *Stenopetalum* (tribe *Microlepidieae*). This study was extended to investigate gene retention after WGD and, most of the *Brassicaceae* mesopolyploid paralogs (including *Pachycladon*) after the At-α event contained similar patterns of gene retention as that of *A. thaliana* in terms of GO categories. Because of this convergent pattern of parallel

retention and loss of functionally related genes after WGD, it has been suggested that the dosage balance hypothesis is playing a role here rather than fates such as sub-functionalization and neo-functionalization.

Utilizing this unique allopolyploid system to explore the fate of duplicated genes in the context of a well-annotated genetic pathway is an exciting prospect. Highly divergent parental genomes, radiation generating multiple species with high morphological diversity and habitat preferences, and a close relationship with the model plant *Arabidopsis thaliana* have all made *Pachycladon* an ideal system to this end. Moreover, we take advantage of the *Arabidopsis* trichome initiation pathway's components being well-described.

1.4 Sequencing approaches to generate gene pathway data

The initial phase of the research described here is involved with the isolation and sequencing of two presumed homeologs of each core gene in the trichome initiation pathway.

1.4.1 First Generation Sequencing - Sanger Sequencing

Sanger sequencing uses the chain termination method using fluorescently tagged DNA polymerase inhibitor 2', 3'-ddNTPs (dideoxyribonucleotide triphosphates) which irreversibly terminate DNA polymerase activity, mixed with natural dNTPs used to synthesize the complementary strand (Metzker 2005). The appropriate mix of these two types of dNTPs generates competition between synthesis and termination resulting in nested fragments which were first analyzed in high-resolution gel electrophoresis to determine the DNA sequence by Frederick Sanger and colleagues in 1977 (Sanger et al 1977). Sanger sequencing was the primary method of DNA sequencing for decades.

As the Sanger method sequences individual DNA fragments at a time, it is ideal for sequencing small regions of DNA from a low number of targets belonging to a limited number of samples (Illumina 2022). It also requires prior sequence information to design gene specific primers in order to use in the sequencing reactions. Generally, 500-700 bp long reads are generated by Sanger sequencing. Sanger sequencing is cost-effective for a small number of

fragments or samples and has a fast workflow without complicated sequence filtering and data analysis, based only on the experimental conditions discussed above. When research interests lie on sequencing single genes, Sanger sequencing is still the most reliable sequencing method with minimal artefact mutations (Gurson 2015).

1.4.2 Next Generation Sequencing (NGS) Technology

Next Generation Sequencing has surpassed the first-generation Sanger method due to its high-throughput sequencing performed at unprecedented speed at rather inexpensive cost when considering large numbers of samples or sequences. NGS sequencing is massively parallel, generating millions of sequences simultaneously per run, thus enabling the sequencing of whole genomes starting with a very low amount of input template. Hence, sequencing volume, depth and sensitivity are much greater compared to Sanger sequencing. This enables greater discovery power to detect rare and novel variants (Gurson 2015).

In contrast to Sanger sequencing NGS generates shorter reads ~150-300 bp, therefore it remains challenging to distinguish repetitive gene sequences when assembling the short sequences together. Therefore, Sanger sequencing is often used to validate the gene sequences obtained from NGS (GenScript 2021).

Generally, NGS sequencing technology has a procedure that consists of template preparation, sequencing, imaging, and data analysis (Metzker 2010). Several NGS technologies have emerged, such as the pyrosequencing method by 454 Life Sciences (Now Roche) (Margulies et al 2005, van Dijk et al 2014), Solexa/Illumina sequencing platform, Sequencing by Oligo Ligation Detection (SOliD) by Applied Biosystems (Now Life Technologies), Qiagen intelligent bio-systems sequencing-by-synthesis (Al-Shehbaz et al 2006, van Dijk et al 2014), Polony sequencing (Shendure et al 2005, van Dijk et al 2014), and a single molecule detection system (Helicos BioSciences) (Pushkarev et al 2009, van Dijk et al 2014). Moreover, the third-generation sequencing technologies are progressing rapidly including Pacific Biosciences (PacBio) and MinION sequencers based on nanopore technology (Lu et al 2016, Munroe & Harris 2010).

Important factors arise when choosing the most appropriate NGS platforms (Mardis 2008, Wolf 2013). Such factors include read length, sequence coverage, price per base pair, error rate, error profile etc. Hence, different platforms generate different forms of data and this data should pave the way to the required form of experimental interpretation at the end (Chu & Corey 2012, Hu et al 2021).

PacBio and Oxford nanopore technology give longer sequencing read lengths, but Illumina has a tremendous increase in throughput and the lowest per-base cost (Illumina HiSeq is generating an astonishing capacity of 1.8 Tb of sequence per run) (Hu et al 2021, van Dijk et al 2014). The current study uses Illumina MiSeq, the most utilized sequencing platform worldwide, which is performed in the benchtop sequencer of Illumina, optimized for low cost and less time input. This easier-to-use diagnostic tool is mainly used for clinical and small laboratory purposes, and it is capable of producing 25 million of 2×300 bp reads (Schirmer et al 2015, van Dijk et al 2014). Moreover, the Illumina technology is compatible with many library preparation protocols and other applications (Liu et al 2012, van Dijk et al 2014).

Library preparation is an elementary step for every NGS workflow. The construction of template for library preparation is made up of a DNA sample which is fragmented first (~50-500 bp) and then selected for the desired size and ligated with adapter sequences. Different adapters are added to each end of fragments, which contain sequences complementary to platform-specific-oligos. They provide read 1 and read 2 universal priming sites for single-end-sequencing or paired-end-sequencing, and also index sequences for bioinformatically sorting out sequencing reads in multiplexed libraries (Metzker 2010, van Dijk et al 2014). Then the template with adapters is amplified to proceed to the sequencing reaction. There are drawbacks in library preparation such as the introduction of quantitative bias and loss of material occurring in fragmentation, end-polishing, adapter ligation, and purification (Goodwin et al 2016, van Dijk et al 2014).

Once the templates are loaded onto the flow cell, subsequent solid-phase bridge amplification occurs to generate clusters of copies of the template. Then, the sequencing-by-synthesis technology uses fluorescently labeled terminator-bound dNTPs for amplification.

These fluorophores are illuminated by a laser and processed for imaging and base calling (Schirmer et al 2015). Finally, the process leads to data analysis after aligning the sequences with reference genome if available or to a *de novo* assembly. The data can be analyzed on the MiSeq instrument, through the web or onsite servers.

1.5 Gene expression analysis

Studying differential gene expression in duplicated gene homeologs is essential to fully understand the effect of whole genome duplication on their transcriptional behavior. Several methods have been developed for gene expression analysis including Northern blot, quantitative reverse transcription PCR (qRT-PCR), serial analysis of gene expression (SAGE), DNA microarrays, and RNA sequencing (RNA-seq) (Git et al 2010, Teo et al 2016). qRT-PCR is ideal for measuring the gene expression of a relatively small number of genes (when the gene of interest or pathway is known) from a large number of samples, whereas RNA-seq is preferred and frequently used for whole transcriptomic expression studies. Although, currently RNA-seq tends to be preferred over microarray due to wider coverage of expression values (Git et al 2010, Wolf 2013).

1.5.1 qRT-PCR (Quantitative Reverse Transcription-PCR)

qRT-PCR is often considered the ‘gold standard’ in the quantification of gene expression, adapting many stepping stones from its development more than two decades ago. The invention of the polymerase chain reaction (PCR) technology by Kary Mullis in 1984 is considered one of the greatest scientific achievements of the 20th century which revolutionized many areas of science including its adaptation to real-time PCR (Ahrberg & Neužil 2015).

A small amount of RNA is adequate for qRT-PCR, which is converted and amplified into cDNA by reverse transcription reaction. The cDNA is used as the template in the PCR reaction and importantly, a fluorescent marker is added to the PCR mixture. Similar to conventional PCR, the number of short specific sections of DNA (delineated by sequence-specific primers) is doubled in every cycle of the qPCR, leading to exponential amplification of targets. Since the amount of fluorescence released during amplification is directly proportional

to the amount of amplified DNA, the qRT-PCR program simultaneously quantifies the amplification of cDNA throughout the PCR during each cycle (Higuchi et al 1993, VanGuilder et al 2008).

A sample amplification plot (Figure 1.6) can be used to understand how qPCR works. The amplification plot shows two main phases, an exponential phase followed by a nonexponential plateau phase. During the initial cycles, the fluorescence emitted by the reporter (probes or dye) is weak due to the small amount of PCR product, and therefore the signal cannot be distinguished from the background signal. Then, the product starts to accumulate exponentially with abundant enzymes and reagents and so the signal increases exponentially, during which time it generates a detectable fluorescent signal over the background noise. The number of cycles required for the sample to reach this threshold level is called the CT value. The higher the initial number of cDNA molecules/sequence-specific oligonucleotide targets in the sample, the faster the fluorescence will increase during the qPCR cycles, which crosses the CT threshold in earlier cycles. If the gene copy numbers are less in the sample, amplification is also reduced with qPCR cycles emitting less fluorescence, which crosses the threshold CT in later cycles. Eventually, with limited reagents the product accumulation decreases, and then the product accumulation ceases in the plateau phase (Kubista et al 2006, VanGuilder et al 2008). A real-time thermocycler detects the amount of DNA amplicons by measuring the fluorescence

Figure 1.6: Sample qPCR amplification plot-a threshold level is set where the sample produces a detectable fluorescence signal over the background noise, and CT indicates the cycle number at which this occurs (Kubista et al 2006).

emitted by probes or dyes. Different fluorescence-based technologies have been developed to use the real-time PCR, such as methods using probe sequences that fluoresce upon hydrolysis (TaqMan-Applied Biosystems) or hybridization (LightCycler-Roche), fluorescent hairpins (LUX- Invitrogen) and intercalating dyes (SYBR Green) (VanGuilder et al 2008).

SYBR Green contains an intercalating dye that fluoresces upon binding to double-stranded DNA. Referring to recent research based on differential gene expression studies using the LightCycler instrument, it seems that the primarily employed chemistry is SYBR GREEN I (Chen et al 2019, Hu et al 2019, Méndez-Vigo et al 2019, Nuruzzaman et al 2008). SYBR GREEN I is a technology that obviates the need for target-specific fluorescent probes, so it is the cheapest option, but it demands strong gene-specific primers to prevent the detection of other nonspecific amplicons. The primers should not form an appreciable amount of primer dimer bands. Further, it is essential to use melting curve analysis for each product to ensure that the fluorescent signal observed is from the desired PCR product, but not from any other nonspecific product (Bustin 2000, Ginzinger 2002).

Experts in the qRT-PCR platform have published a set of guidelines to help ensure the reliability, reproducibility, and transparency of qPCR results between laboratories (Bustin et al 2009). This set of guidelines is named the MIQE guidelines (Minimum Information for Publication of Quantitative Real-Time PCR Experiments) (Bustin et al 2009), which includes information such as experimental design, sample acquisition, handling, and preparation, QC (Quality Control) of nucleic acids, reverse transcription, qPCR target information, qPCR oligonucleotides, qPCR protocol, normalization, validation, and data analysis. Table A1 included in the appendix contains a complete MIQE checklist including all the information required for successfully reporting a qPCR study.

It is essential to normalize qPCR assays to correct the apparent variations that arise from different samples' extraction yields, reverse transcription yields, and amplification efficiencies. Thus, normalization enables the comparison of mRNAs across different samples. The use of reference genes, sometimes referred to as "housekeeping genes" is the most

common method of normalizing cellular mRNA data. Ideally, reference genes are expected to be stably expressed across different tissue types and different samples, and strongly correlated by abundance to the total amounts of mRNA present in samples (Bustin et al 2009, Hruz et al 2011). For plants, there are traditional reference genes that have frequently been used, such as beta-actin (ACTB), ubiquitin (UBQ), or glyceraldehyde 3-phosphate dehydrogenase (GAPDH) (Hruz et al 2011). However, it is strongly recommended to validate the suitability of available reference genes for every new experiment, as their expression could vary considerably depending on specific experimental conditions. Generally, it is more acceptable to use multiple reference genes rather than a single one (unless with a clear justification of invariant expression of the single reference gene under the experiment conditions being tested) to improve the accuracy, quality and sensitivity of the experiment. Statistical models are improved to select the best reference genes and number of genes when there are multiple reference genes. Three of the most commonly used algorithms are GeNorm (Vandesompele et al 2002), Norm finder (Andersen et al 2004), and Bestkeeper (Pfaffl et al 2004).

Real-time qPCR data analysis can be performed in two quantification strategies: either absolute or relative quantification. Absolute quantification measures the copy number of the product of interest, using a calibration curve with known concentrations of gene of interest. This results in an amount of gene copy number per given amount of mRNA sample. Relative quantification measures the expression level of the gene of interest relative to the reference gene(s). The result is a fold difference of gene expression compared to the control sample; e.g. fold difference in expression of a chemically treated sample compared to untreated control sample (Pfaffl 2012).

1.6 Thesis structure and objectives

The range of gene fates following duplication is theoretically well described; however, little is known about how networks of interacting genes and proteins respond to wholesale duplication, despite that this is theoretically a key determinant in the fate of duplicated genes. We have identified a conserved trichome initiation genetic pathway (in the *Brassicaceae*) that

regulates several morphological and biochemical phenotypes, as a good candidate system for studying duplicated genetic pathways. We have further identified an ideal study group, *Pachycladon*, that has an evolutionary history of allopolyploidy that would have initially possessed quite diverged copies of the genetic pathway of interest.

Chapter 2

Chapter 2 describes the content related to the following objective 1.6.1, on sequencing duplicate gene copies and placing them in phylogenetically to deduce which copy of each gene (homeolog) belongs to each ancestral pathway.

1.6.1 Reconstruct and characterize the "ancestral" epidermal cell fate pathways from multiple *Pachycladon* species, with a focus on the trichome initiation pathway

1.6.1.1 Isolate and sequence both homeologs (where present) of the primary genes that direct epidermal cell fate

The initial objective is to assemble and analyse the "ancestral" genetic pathways (homeologs of each gene in the pathway) from four species of the allopolyploid genus, *Pachycladon*, that represent the phylogenetic diversity of the genus (*P. enysii*, *P. fastigiatum*, *P. novae-zealandiae*, and *P. cheesemanii*). Preliminary data have shown that for each gene of interest, there are two strongly diverged copies in *Pachycladon* species, representing the ancestral orthologs. To isolate both homeologs of each gene of interest, a combination of iterative PCR and Sanger sequencing and MiSeq genomic and cDNA library skim-sequencing attempts were taken. Complete coding sequences for homeologs were attempted to assemble as much as possible from multiple species, and the sequences were analyzed in multiple ways in the following chapters, as described below. The genes of interest consist of *TTG1*, *GL1*, *GL3*, and *EGL3*, the core genes that form the primary trichome initiation complex.

1.6.1.2 Reconstruct and contrast the two progenitor trichome development pathways by phylogenetic analysis

The full-length coding sequences from all homeologs of a given gene are analyzed with full-length orthologs from species representing as much diversity of the *Brassicaceae* as possible, and with outgroup sequences from species in the *Cleomaceae*, where possible. Following the alignment of these sequences for a given gene, a phylogenetic analysis is performed, and the resulting trees are rooted with the outgroup sequence. The objective here is to identify the phylogenetic placement of each of the *Pachycladon* homeologs for a given gene and to compare different gene trees to identify which copy of each gene goes with another to group the copies involved in each parental subgenome. We hypothesize that two distinct orthologs of each gene of interest will be isolated from each *Pachycladon* species and that similar patterns of phylogenetic placement for the two copies of each gene will be evident; i.e., in each case, one copy will cluster with orthologs from a particular Brassicaceae lineage and the other copy will cluster with a different lineage, but those two lineages will be consistent across genes. Our preliminary data have shown that one homeolog for each gene tends to have a fairly close affiliation with the *A. thaliana* ortholog and the other sits well outside that cluster. Thus, the ancestral (progenitor) species' trichome initiation pathways are reconstructed, after assigning each homeolog to the correct group, allowing for further molecular evolutionary comparisons in the next chapters.

Chapter 3

The sequence data produced in Chapter 2 is analyzed in Chapter 3, to determine how molecular evolution acts when genes are duplicated in an important genetic pathway post-polyploidization.

1.6.2 Analyzing and characterizing the molecular evolution of the trichome initiation pathway across species

The data generated in the second chapter are used for some analyses such as (1) homeolog sequences presence/absence as indications of duplicate gene loss; (2) quantification of homeolog polymorphism and divergence for each gene of interest; and (3) exploration of the protein structure and function, to characterize the pathway in terms of homeolog loss and retention combined with the molecular attributes such as pleiotropy, epistasis, functional overlap and parental subgenome bias, selection pressures on homeologs, and to infer potential functional divergences of the homeologs. In the protein analysis, structural and functional annotations were largely based on available sequence and functional annotations, particularly from *A. thaliana*.

If one pathway Type belonging to one subgenome is preferentially retained over the other Type belonging to the other subgenome, we predict that all gene/protein copies of the former Type are intact, less diverse, and under purifying selection compared to the gene copies belonging to the latter Type. On the other hand, the pathway may have evolved to function with the combination of interacting homeologs of different Types without a subgenome bias.

Chapter 4

Chapter 4 presents new expression data for homeologs of the genes of interest, to investigate the effect of gene duplication towards differential gene expression of gene copies in a genetic pathway.

1.6.3 Expression analysis of homeologous genes to analyze differential expression

Among the changes that occur following gene duplication is an alteration of expression patterns. With the ancestral pathways reconstructed, we measure the expression levels of both homeologs of each gene of interest using RT-qPCR (LightCycler, Roche). This work is carried out on multiple individuals per species from young or mature developmental stages (for *P. enysii*, *P. fastigiatum*, and *P. novae-zealandiae*) and two phases of development (as young leaves and fully expanded mature leaves for *P. cheesemanii*), to characterize the expression profiles of homeologs across the pathway.

The gene expression results are mainly analyzed to determine if there are any differential expression patterns among the homeologs and *Pachycladon* species related to models such as subgenome-biased expression or redundant expression. Also, the impact of young and mature developmental stages of *P. cheesemanii* on gene expression is investigated. Overall, a comprehensive discussion is presented with the genetic diversity data found and evolutionary assumptions made in the previous chapter.

Here we make predictions that align with earlier prediciton. If one of the subgenomes seems to be retained more than the other, we predict that the former subgenome will have expression bias over the latter subgenome.

Chapter 5

This is the summary of the thesis, drawing together the results of each chapter in a general discussion. The findings of this thesis are related to other studies and future directions for research of gene duplication fate in a genetic pathway context are suggested.

REFERENCES

Adams K, Wendel J. 2004. Exploring the genomic mysteries of polyploidy in cotton. *Biological Journal of the Linnean Society* 82: 573-81

Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168: 2217-26

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8: 135-41

Ahrberg CD, Neužil P. 2015. Doubling Throughput of a Real-Time PCR. *Scientific Reports* 5: 12595

Al-Shehbaz IA, Beilstein MA, Kellogg EA. 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution* 259: 89-120

Al-Shehbaz IA, O'Kane SL, Price RA. 1999. Generic Placement of Species Excluded from Arabidopsis (Brassicaceae). *Novon* 9: 296-307

Alfandary R. 2015. 4 Tips for Efficient Primer Design.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215: 403-10

Andersen CL, Jensen JL, Ørntoft TF. 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer research* 64: 5245-50

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444: 171-78

Balkunde R, Pesch M, Hülskamp M. 2010. Trichome patterning in Arabidopsis thaliana: from genetic to molecular models. *Current topics in developmental biology* 91: 299-321

Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B. 2011. Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. *New Phytol* 192: 760-74

Behling AH, Shepherd LD, Cox MP. 2020. The importance and prevalence of allopolyploidy in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand* 50: 189-210

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American journal of botany* 93: 607-19

Bennett MD. 2004. Perspectives on polyploidy in plants–ancient and neo. *Biological Journal of the Linnean Society* 82: 411-23

Bernhardt C, Lee MM, Gonzalez A, Zhang F, Lloyd A, Schiefelbein J. 2003. The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root. *Development* 130: 6431-39

Bicknell RA, Heenan PB, Dawson MI, Fletcher PJ, Christie MC. 2009. Matromorphy in Pachycladon exile (brassicaceae) revealed by interspecific hybridisation. *New Zealand Journal of Botany* 47: 139-48

Birchler JA. 2012. Genetic consequences of polyploidy in plants In *Polyploidy and genome evolution*, pp. 21-32: Springer

Blakeslee AF. 1941. Effect of induced polyploidy in plants. *The American Naturalist* 75: 117-35

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* 13: 137-44

Blanc G, Wolfe KH. 2004a. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16: 1679-91

Blanc G, Wolfe KH. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* 16: 1667-78

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433-38

Bridges CB. 1936. The bar" gene" a duplication. *Science* 83: 210-11

Buggs RJ, Chamala S, Wu W, Gao L, May GD, et al. 2010. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular ecology* 19: 132-46

Buggs RJ, Chamala S, Wu W, Tate JA, Schnable PS, et al. 2012. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr Biol* 22: 248-52
Bustin SA. 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* 25: 169-93
Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, et al. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55: 611-22
Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martínez-García JF, Bilbao-Castro JR, Robertson DL. 2010. Genome-Wide Classification and Evolutionary Analysis of the bHLH Family of Transcription Factors in Arabidopsis, Poplar, Rice, Moss, and Algae *Plant Physiology* 153: 1398-412
Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biology* 11: R125
Chapman MA, Burke JM. 2007. Genetic divergence and hybrid speciation. *Evolution* 61: 1773-80
Chen G, Li X, Qiao X, Li J, Wang L, et al. 2019. Genome-wide survey and expression analysis of the SLAC/SLAH gene family in pear (Pyrus bretschneideri) and other members of the Rosaceae. *Genomics* 111: 1097-107
Chen S, Wang S. 2019. GLABRA2, a Common Regulator for Epidermal Cell Fate Determination and Anthocyanin Biosynthesis in Arabidopsis. *International Journal of Molecular Sciences* 20: 4997
Chen ZJ. 2007. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual Review of Plant Biology* 58: 377-406
Chen ZJ, Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* 28: 240-52
Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* 4: 258-68
Chezem WR, Clay NK. 2016. Regulation of plant secondary metabolism and associated specialized cell development by MYBs and bHLHs. *Phytochemistry* 131: 26-43
Chu Y, Corey DR. 2012. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther* 22: 271-4
Comai L. 2000. Genetic and epigenetic interactions in allopolyploid plants. *Plant Gene Silencing*: 267-79
Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6: 836-46
Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* 19: 91-98
Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome research* 16: 738-49
D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488: 213-17
De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20: 591-7
De Vries H. 1910. *The mutation theory: Experiments and observations on the origin of species in the vegetable kingdom*. Kegan Paul.
Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research* 12: 1075-79
DeVries H. 1915. The coefficient of mutation in Oenothera biennis L. *Botanical Gazette* 59: 169-96
Díaz-Pérez A, López-Álvarez D, Sancho R, Catalán P. 2018. Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus Brachypodium using minimum evolution, coalescence and maximum likelihood approaches. *Molecular Phylogenetics and Evolution* 127: 256-71
Dodsworth S, Chase MW, Leitch AR. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180: 1-5

Dong Y, Gupta S, Sievers R, Wargent JJ, Wheeler D, et al. 2019. Genome draft of the Arabidopsis relative Pachycladon cheesemanii reveals novel strategies to tolerate New Zealand's high ultraviolet B radiation environment. *BMC Genomics* 20: 838
Doroshkov AV, Konstantinov DK, Afonnikov DA, Gunbin KV. 2019. The evolution of gene regulatory networks controlling Arabidopsis thaliana L. trichome development. *BMC Plant Biology* 19: 53
Doyle JJ, Coate JE. 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180: 1-52
Doyle JJ, Doyle JL. 1990. Isolation ofplant DNA from fresh tissue. *Focus* 12: 39-40
Doyle JJ, Sherman-Broyles S. 2017. Double trouble: taxonomy and definitions of polyploidy. *New Phytologist* 213: 487-93
Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* 17: 699-717
Farhat P, Hidalgo O, Robert T, Siljak-Yakovlev S, Leitch IJ, et al. 2019. Polyploidy in the Conifer Genus Juniperus: An Unexpectedly High Rate. *Front Plant Sci* 10: 676-76
Ferreira de Carvalho J, Lucas J, Deniot G, Falentin C, Filangi O, et al. 2019. Cytonuclear interactions remain stable during allopolyploid evolution despite repeated whole-genome duplications in Brassica. *Plant J* 98: 434-47
Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytol* 183: 557-64
Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML. 2007. Evolutionary dynamics of Waxy and the origin of hexaploid Spartina species (Poaceae). *Mol Phylogenet Evol* 43: 1040-55
Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn* 4: 25-40
Freeling M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology* 60: 433-53
Freeling M, Scanlon MJ, Fowler JE. 2015. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr Opin Genet Dev* 35: 110-8
Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell* 19: 3403-17
Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in Arabidopsis. *Molecular biology and evolution* 24: 2298-309
Garnock-Jones PJ. 1991. Gender dimorphism in Cheesemania wallii (Brassicaceae). *New Zealand Journal of Botany* 29: 87-90
Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC. 2000. Maize as a model for the evolution of plant nuclear genomes. *Proceedings of the National Academy of Sciences* 97: 7008-15
Geneious 9.1.8. 2017.
GenScript. 2021. Sanger Sequencing vs. Next-Generation Sequencing (NGS).
German DA, Friesen N, Neuffer B, Al-Shehbaz IA, Hurka H. 2009. Contribution to ITS phylogeny of the Brassicaceae, with special reference to some Asian taxa. *Plant Systematics and Evolution* 283: 33-56
Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome research* 17: 669-81
Gerstel D, Phillips L. 1957. Segregation in new allopolyploids of Gossypium. II. Tetraploid combinations. *Genetics* 42: 783
Ginzinger DG. 2002. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. *Experimental hematology* 30: 503-12
Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, et al. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *Rna* 16: 991-1006
Gong L, Olson M, Wendel JF. 2014. Cytonuclear Evolution of Rubisco in Four Allopolyploid Lineages. *Molecular Biology and Evolution* 31: 2624-36

Gong L, Salmon A, Yoo M-J, Grupp KK, Wang Z, et al. 2012. The Cytonuclear Dimension of Allopolyploid Evolution: An Example from Cotton Using Rubisco. *Molecular Biology and Evolution* 29: 3023-36
Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17: 333-51
Grant V. 1981. Plant speciation, Columbia University Press
Grebe M. 2012. The patterning of epidermal hairs in Arabidopsis--updated. *Curr Opin Plant Biol* 15: 31-7
Grover C, Gallagher J, Szadkowski E, Yoo M, Flagel L, Wendel J. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* 196: 966-71
Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, et al. 2015. Re-evaluating the phylogeny of allopolyploid Gossypium L. *Molecular Phylogenetics and Evolution* 92: 45-52
Gurson N. 2015. When Do I Use Sanger Sequencing vs NGS? : ThermoFisher Scientific
Han G, Li Y, Yang Z, Wang C, Zhang Y, Wang B. 2022. Molecular mechanisms of plant trichome development. *Front Plant Sci* 13: 910228
Hao Y, Zong X, Ren P, Qian Y, Fu A. 2021. Basic Helix-Loop-Helix (bHLH) transcription factors regulate a wide range of functions in Arabidopsis. *International Journal of Molecular Sciences* 22: 7152
Hauser M-T. 2014. Molecular basis of natural variation and environmental control of trichome patterning. *Front Plant Sci* 5
Heenan P, Mitchell A, Koch M. 2002. Molecular systematics of the New Zealand Pachycladon (Brassicaceae) complex: generic circumscription and relationships to Arabidopsis sens. lat. and Arabis sens. lat. *New Zealand Journal of Botany* 40: 543-62
Heenan PB. 1999. Artificial intergeneric hybrids between the New Zealand endemic Ischnocarpus and Pachycladon (Brassicaceae). *New Zealand Journal of Botany* 37: 595-601
Heenan PB, Dawson MI, Smissen RD, Bicknell RA. 2008. An artificial intergeneric hybrid derived from sexual hybridization between the distantly related Arabidopsis thaliana and Pachycladon cheesemanii (Brassicaceae). *Botanical Journal of the Linnean Society* 157: 533-44
Heenan PB, Mitchell AD. 2003. Phylogeny, biogeography and adaptive radiation of *Pachycladon* (Brassicaceae) in the mountains of South Island, New Zealand. *Journal of Biogeography* 30: 1737–49
Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. 2006. Transcriptome Shock after Interspecific Hybridization in Senecio Is Ameliorated by Genome Duplication. *Current Biology* 16: 1652-59
Hendriks KP, Kiefer C, Al-Shehbaz IA, Bailey CD, van Huysduynen AH, et al. 2023. Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Current Biology*
Higuchi R, Fockler C, Dollinger G, Watson R. 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology* 11: 1026-30
Hruz T, Wyss M, Docquier M, Pfaffl MW, Masanetz S, et al. 2011. RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* 12: 156
Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: An overview. *Human Immunology* 82: 801-11
Hu X, Hao C, Cheng Z-M, Zhong Y. 2019. Genome-Wide Identification, Characterization, and Expression Analysis of the Grapevine Superoxide Dismutase (SOD) Family. *International Journal of Genomics* 2019: 13
Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, et al. 2001. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* 29: 102-5
Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, et al. 2015. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. *Molecular Biology and Evolution* 33: 394-412
Hufton AL, Panopoulou G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Current opinion in genetics & development* 19: 600-06

Hülskamp M, Misŕa S, Jürgens G. 1994. Genetic dissection of trichome cell development in Arabidopsis. *Cell* 76: 555-66
Illumina. 2022. Key differences between next-generation sequencing and Sanger sequencing-Understanding when NGS can be a more effective option.
Ishida T, Kurata T, Okada K, Wada T. 2008. A genetic regulatory network in the development of trichomes and root hairs. *Annu. Rev. Plant Biol.* 59: 365-86
Jiang W-k, Liu Y-l, Xia E-h, Gao L-z. 2013. Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants *Plant Physiology* 161: 1844-61
Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97
Johnson CS, Kolevski B, Smyth DR. 2002. TRANSPARENT TESTA GLABRA2, a Trichome and Seed Coat Development Gene of Arabidopsis, Encodes a WRKY Transcription Factor. *The Plant Cell Online* 14: 1359-75
Joly S, Heenan PB, Lockhart PJ. 2009. A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. *Mol Phylogenet Evol* 51: 365-72
Joly S, Heenan PB, Lockhart PJ. 2013. Species Radiation by Niche Shifts in New Zealand's Rockcresses (*Pachycladon*, Brassicaceae). *Syst. Biol.* 63: 192–202
Joly S, Heenan PB, Lockhart PJ. 2014. Species Radiation by Niche Shifts in New Zealand's Rockcresses (Pachycladon, Brassicaceae). *Systematic Biology* 63: 192-202
Joly S, Rauscher JT, Sherman-Broyles SL, Brown AH, Doyle JJ. 2004. Evolutionary dynamics and preferential expression of homeologous 18S-5.8S-26S nuclear ribosomal genes in natural and artificial glycine allopolyploids. *Mol Biol Evol* 21: 1409-21
Joshi P, Ansari H, Dickson R, Ellison NW, Skema C, Tate JA. 2022. Polyploidy on islands – concerted evolution and gene loss amid chromosomal stasis. *Annals of Botany* 131: 33-44
Kaltenegger E, Leng S, Heyl A. 2018. The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. *BMC Evol Biol* 18: 76
Karabourniotis G, Liakopoulos G, Nikolopoulos D, Bresta P. 2020. Protective and defensive roles of non-glandular trichomes against multiple stresses: structure–function coordination. *Journal of Forestry Research* 31: 1-12
Kihara H, Ono T. 1926. Chromosomenzahlen und systematische gruppierung der Rumex-arten. *Zeitschrift für Zellforschung und mikroskopische Anatomie* 4: 475-81
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* 279: 5048-57
Kryvych S, Kleessen S, Ebert B, Kersten B, Fisahn J. 2011. Proteomics–the key to understanding systems biology of Arabidopsis trichomes. *Phytochemistry* 72: 1061-70
Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, et al. 2006. The real-time polymerase chain reaction. *Molecular Aspects of Medicine* 27: 95-125
Kurata T, Ishida T, Kawabata-Awai C, Noguchi M, Hattori S, et al. 2005. Cell-to-cell movement of the CAPRICE protein in Arabidopsis root epidermal cell differentiation. *Development* 132: 5387-98
Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. 2004. Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics* 166: 935-45
Larkin JC, Oppenheimer DG, Lloyd AM, Paparozzi ET, Marks MD. 1994. Roles of the GLABROUS1 and TRANSPARENT TESTA GLABRA Genes in Arabidopsis Trichome Development. *The Plant cell* 6: 1065-76
Larkin JC, Walker JD, Bolognesi-Winfield AC, Gray JC, Walker AR. 1999. Allele-specific interactions between ttg and gl1 during trichome development in Arabidopsis thaliana. *Genetics* 151: 1591-604
Lee S, Choi S, Jeon D, Kang Y, Kim C. 2020. Evolutionary impact of whole genome duplication in Poaceae family. *Journal of Crop Science and Biotechnology* 23: 413-25

Levy AA, Feldman M. 2004. Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biological journal of the Linnean society* 82: 607-13
Li X, Noll M. 1994. Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions. *Nature* 367: 83-87
Li Z, McKibben MT, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72: 387-410
Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, et al. 2008. Rapid chromosome evolution in recently formed polyploids in Tragopogon (Asteraceae). *PloS one* 3: e3353
Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol* 29: 365-79
Liu B, Xu C, Zhao N, Qi B, Kimatu JN, et al. 2009. Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. *Journal of Genetics and Genomics* 36: 519-28
Liu C, Zhang TZ. 2019. Functional diversifications of GhERF1 duplicate genes after the formation of allotetraploid cotton. *J Integr Plant Biol* 61: 60-74
Liu L, Li Y, Li S, Hu N, He Y, et al. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012: 251364
Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* 14: 265-79
Lu Y, Rausher MD. 2003. Evolutionary Rate Variation in Anthocyanin Pathway Genes. *Molecular Biology and Evolution* 20: 1844-53
Luo X, Hu Q, Zhou P, Zhang D, Wang Q, et al. 2017. Chasing ghosts: allopolyploid origin of Oxyria sinensis (Polygonaceae) from its only diploid congener and an unknown ancestor. *Molecular Ecology* 26: 3037-49
Mable B. 2004. 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society* 82: 453-66
Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, et al. 2005. Genomic changes in synthetic Arabidopsis polyploids. *The Plant Journal* 41: 221-30
Maes L, Inzé D, Goossens A. 2008. Functional Specialization of the TRANSPARENT TESTA GLABRA1 Network Allows Differential Hormonal Control of Laminal and Marginal Trichome Initiation in Arabidopsis Rosette Leaves. *Plant Physiology* 148: 1453-64
Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. *Journal of Genetics* 92: 155-61
Manda´kova´ T, Joly S, Krzywinski M, Mummenhoff K, Lysaka MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell* 22: 2277–90
Mandáková T, Heenan PB, Lysak MA. 2010. Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evol Biol* 10: 367-67
Mandakova T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. *Plant Cell* 22: 2277-90
Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal* 91: 3-21
Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus Viola (Violaceae). *Systematic biology* 64: 84-101
Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133-41
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-80
Markus Lange B, Turner GW. 2013. Terpenoid biosynthesis in trichomes—current status and future opportunities. *Plant biotechnology journal* 11: 2-22
McBreen K, Heenan PB. 2006. Phylogenetic relationships of Pachycladon (Brassicaceae) species based on three nuclear and two chloroplast DNA markers. *New Zealand Journal of Botany* 44: 377-86
McClintock B. 1929. A cytological and genetical study of triploid maize. *Genetics* 14: 180

McCollum GD. 1957. Comparative studies of chromosome pairing in natural and induced tetraploid Dactylis. *Chromosoma* 9: 571-605
Méndez-Vigo B, Ausín I, Zhu W, Mollá-Morales A, Balasubramanian S, Alonso-Blanco C. 2019. Genetic Interactions and Molecular Evolution of the Duplicated Genes ICARUS2 and ICARUS1 Help Arabidopsis Plants Adapt to Different Ambient Temperatures. *The Plant Cell* 31: 1222-37
Metzker ML. 2005. Emerging technologies in DNA sequencing. *Genome Res* 15: 1767-76
Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46
Mirzaei M, Pascovici D, Keighley T, George I, Voelckel C, et al. 2011. Shotgun proteomic profiling of five species of New Zealand Pachycladon. *Proteomics* 11: 166-71
Mitchell AD, Heenan PB. 2000. Systematic Relationships of New Zealand Endemic Brassicaceae Inferred from nrDNA ITS Sequence Data. *Systematic Botany* 25: 98-105
Mitchell AD, Heenan PB. 2002. Genetic variation within the Pachycladon (Brassicaceae) complex based on fluorescent AFLP data. *Journal of the Royal Society of New Zealand* 32: 427-43
Moghe GD, Shiu SH. 2014. The causes and molecular consequences of polyploidy in flowering plants. *Ann N Y Acad Sci* 1320: 16-34
Morohashi K, Grotewold E. 2009. A Systems Approach Reveals Regulatory Circuitry for Arabidopsis Trichome Initiation by the GL3 and GL1 Selectors. *PLOS Genetics* 5: e1000396
Morrison J, Rajhathy T. 1960. Frequency of quadrivalents in autotetraploid plants. *Nature* 187: 528-30
Munroe DJ, Harris TJR. 2010. Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology* 28: 426-28
Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, USA.
Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, et al. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222: 1638-51
Nuruzzaman M, Gupta M, Zhang C, Wang L, Xie W, et al. 2008. Sequence and expression analysis of the thioredoxin protein gene family in rice. *Molecular Genetics and Genomics* 280: 139
Ohno S. 1970. *Evolution by Gene Duplication*. Springer.
Ohno S. 2013. *Evolution by gene duplication*. Springer Science & Business Media.
Okamoto S, Negishi K, Toyama Y, Ushijima T, Morohashi K. 2020. Leaf Trichome Distribution Pattern in Arabidopsis Reveals Gene Expression Variation Associated with Environmental Adaptation. *Plants* 9: 909
Ownbey M, McCollum GD. 1953. Cytoplasmic inheritance and reciprocal amphiploidy in Tragopogon. *American journal of botany* 40: 788-96
Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of Gene Duplication in Plants. *Plant Physiology* 171: 2294-316
Paterson A, Wang X, Tang H, Lee T. 2012. *Synteny and genomic rearrangements*. Springer.
Pattanaik S, Patra B, Singh SK, Yuan L. 2014. An overview of the gene regulatory network controlling trichome development in the model plant, Arabidopsis. *Front Plant Sci* 5: 259
Payne CT, Zhang F, Lloyd AM. 2000. GL3 Encodes a bHLH Protein That Regulates Trichome Development in Arabidopsis Through Interaction With GL1 and TTG1. *Genetics* 156: 1349-62
Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes* 9: 88
Pfaffl MW. 2012. Quantification strategies in real-time polymerase chain reaction. *Quantitative real-time PCR Appl Microbiol*: 53-62
Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology letters* 26: 509-15
Pires JC, Zhao J, Schranz ME, Leon EJ, Quijada PA, et al. 2004. Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). *Biological Journal of the Linnean Society* 82: 675-88
Price JP, Wagner WL. 2004. Speciation in Hawaiian Angiosperm Lineages: Cause, Consequence, and Mode. *Evolution* 58: 2185-200
Price RA, Al-Shehbaz IA, O'Kane SL. 2001. Beringia (Brassicaceae), a New Genus of Arabidopsoid Affinities from Russia and North America. *Novon* 11: 332-36

Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. *Nature Biotechnology* 27: 847
Qi B, Huang W, Zhu B, Zhong X, Guo J, et al. 2012. Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. *BMC Biology* 10: 3
Ramsey J, Ramsey TS. 2014. Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130352
Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics* 29: 467-501
Ramsey J, Schemske DW. 2002. Neopolyploidy in Flowering Plants. *Annual Review of Ecology and Systematics* 33: 589-639
Rapp RA, Udall JA, Wendel JF. 2009. Genomic expression dominance in allopolyploids. *BMC Biology* 7: 18
Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol* 16: 266-74
Rinehart JA, Petersen MW, John ME. 1996. Tissue-specific and developmental regulation of cotton gene FbL2A (demonstration of promoter activity in transgenic plants). *Plant Physiology* 112: 1331-41
Román-Palacios C, Medina CA, Zhan SH, Barker MS. 2021. Animal chromosome counts reveal a similar range of chromosome numbers but with less polyploidy in animals compared to flowering plants. *Journal of Evolutionary Biology* 34: 1333-39
Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, et al. 2013. The fate of duplicated genes in a polyploid plant genome. *The Plant Journal* 73: 143-53
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci. USA* 74: 5463-67
Satina S, Blakeslee AF. 1937. Chromosome behavior in triploids of Datura stramonium. I. The male gametophyte. *American journal of botany*: 518-27
Sattler MC, Carvalho CR, Clarindo WR. 2016. The polyploidy and its key role in plant breeding. *Planta* 243: 281-96
Schenk JJ. 2021. The Next Generation of Adaptive Radiation Studies in Plants. *International Journal of Plant Sciences* 182: 245-62
Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* 43: e37-e37
Schlueter JA, Dixon P, Granger C, Grant D, Clark L, et al. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868-76
Schranz ME, Mitchell-Olds T. 2006. Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae. *The Plant Cell* 18: 1152-65
Schwab B, Folkers U, Ilgenfritz H, Hülskamp M. 2000. Trichome morphogenesis in Arabidopsis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355: 879-83
Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, et al. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* 20: 633-43
Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309: 1728-32
Shi X, Zhang C, Ko DK, Chen ZJ. 2015. Genome-Wide Dosage-Dependent and -Independent Regulation Contributes to Gene Expression and Evolutionary Novelty in Plant Polyploids. *Molecular Biology and Evolution* 32: 2351-66
Shimeld SM. 1999. Gene function, gene networks and the fate of duplicated genes. *Seminars in Cell & Developmental Biology* 10: 549-53
Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 99: 13627-32

Singh ND, Kumar S, Daniell H. 2016. Expression of β-glucosidase increases trichome density and artemisinin content in transgenic Artemisia annua plants. *Plant biotechnology journal* 14: 1034-45
Soltis DE, Buggs RJ, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* 59: 1387-403
Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in ecology & evolution* 14: 348-52
Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant speciation. *New Phytologist* 161: 173-91
Soltis DE, Soltis PS, Tate JA. 2004. Advances in the study of polyploidy since Plant speciation. *New Phytologist* 161: 173-91
Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then...and now: Stebbins revisited. *American journal of botany* 101: 1057-78
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current opinion in genetics & development* 35: 119-25
Song K, Lu P, Tang K, Osborn TC. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* 92: 7719-23
Stebbins Jr CL. 1950. Variation and evolution in plants. *Variation and evolution in plants.*
Stuessy TF, Crawford DJ. 1998. Chromosomal stasis during speciation in angiosperms of oceanic islands In *Evolution and Speciation of Island Plants*, ed. M Ono, TF Stuessy, pp. 307-24. Cambridge: Cambridge University Press
Sumitomo K, Shirasawa K, Isobe S, Hirakawa H, Hisamatsu T, et al. 2019. Genome-wide association study overcomes the genome complexity in autohexaploid chrysanthemum and tags SNP markers onto the flower color genes. *Scientific Reports* 9: 13947
Symonds VV, Hatlestad G, Lloyd AM. 2011. Natural allelic variation defines a role for ATMYC1: trichome cell fate determination. *PLoS Genet* 7: e1002069
Tamayo-Ordóñez M, Espinosa-Barrera L, Tamayo-Ordóñez Y, Ayil-Gutiérrez B, Sánchez-Teyer L. 2016. Advances and perspectives in the generation of polyploid plant species. *Euphytica* 209: 1-22
Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38: 3022-27
Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, et al. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics* 190: 1563-74
TATE JA, SOLTIS DE, SOLTIS PS. 2005. Polyploidy in plants In *The evolution of the genome*, pp. 371-426: Elsevier
Teo ZL, Savas P, Loi S. 2016. Gene Expression Analysis: Current Methods In *Molecular Pathology in Cancer Research*, ed. SR Lakhani, SB Fox, pp. 107-36. New York, NY: Springer New York
Toledo-Ortiz G, Huq E, Quail PH. 2003. The Arabidopsis basic/helix-loop-helix transcription factor family. *The Plant Cell* 15: 1749-70
Tominaga-Wada R, Ishida T, Wada T. 2011. New insights into the mechanism of development of Arabidopsis root hairs and trichomes. *International review of cell and molecular biology* 286: 67-106
Udall JA, Quijada PA, Osborn TC. 2005. Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of Brassica napus L. *Genetics* 169: 967-79
Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* 35: W71-W74
Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* 33: 11-26
Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009. The flowering world: a tale of duplications. *Trends in Plant Science* 14: 680-88
Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411-24

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* 30: 418-26
Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, et al. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology* 3: 1-12
VanGuilder HD, Vrana KE, Freeman WM. 2008. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* 44: 619-26
Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. *Science* 290: 2114-17
Vries Hd. 1901. Die Mutationstheorie: Versuche und Beobachtungen über die Entstehung von Arten im Pflanzenreich. *(No Title)*
Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N, et al. 1999. The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. *The Plant cell* 11: 1337-50
Wang J, Roe B, Macmil S, Yu Q, Murray JE, et al. 2010. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11: 261
Wang J, Tian L, Madlung A, Lee HS, Chen M, et al. 2004. Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics* 167: 1961-73
Wang L, Si W, Yao Y, Tian D, Araki H, Yang S. 2012. Genome-wide survey of pseudogenes in 80 fully re-sequenced Arabidopsis thaliana accessions. *PloS one* 7: e51769
Wang S, Chen J-G. 2014. Regulation of cell fate determination by single-repeat R3 MYB transcription factors in Arabidopsis. *Front Plant Sci* 5
Wang X, Morton JA, Pellicer J, Leitch IJ, Leitch AR. 2021a. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal* 107: 1003-15
Wang X, Shen C, Meng P, Tan G, Lv L. 2021b. Analysis and review of trichomes in plants. *BMC plant biology* 21: 1-11
Wang X, Wang H, Wang J, Sun R, Wu J, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* 43: 1035-39
Wang X, Zhang N, Zhao Y, Wang J. 2021c. A New Method for Recognizing Protein Complexes Based on Protein Interaction Networks and GO Terms. *Front Genet* 12: 792265
Wang Z, Yang Z, Li F. 2019. Updates on molecular mechanisms in the development of branched trichome in Arabidopsis and nonbranched in cotton. *Plant Biotechnology Journal* 17: 1706-22
Weiss-Schneeweiss H, Emadzade K, Jang T-S, Schneeweiss GM. 2013. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenetic and genome research* 140: 137-50
Wendel JF. 2000. Genome evolution in polyploids. *Plant molecular evolution*: 225-49
Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr Opin Genet Dev* 49: 1-7
Winkler H. 1916. Uber die experimentelle Erzeugung von Pflanzen mit abweichen den Chromosomenzahlen. Zeitschr. f. *Bot. Bd* 8
Wolf JB. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol Ecol Resour* 13: 559-72
Xiong Z, Gaeta RT, Pires JC. 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. *Proceedings of the National Academy of Sciences* 108: 7908-13
Yang C, Ye Z. 2013. Trichomes as models for studying plant cell differentiation. *Cellular and molecular life sciences* 70: 1937-48
Yogeeswaran K, Voelckel C, Joly S, Heenan PB. 2011. Pachycladon. 227-49
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18: 821-29
Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. 2003. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. *Development* 130: 4859-69
Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18: 292-98
Zhao B, Liu L, Tan D, Wang J. 2010. Analysis of phylogenetic relationships of Brassicaceae species based on Chs sequences. *Biochemical Systematics and Ecology* 38: 731-39

Zhao H, Li X, Ma L. 2012a. Basic helix-loop-helix transcription factors and epidermal cell fate determination in Arabidopsis. *Plant Signal Behav* 7: 1556-60

Zhao H, Wang X, Zhu D, Cui S, Li X, et al. 2012b. A Single Amino Acid Substitution in IIIf Subfamily of Basic Helix-Loop-Helix Transcription Factor AtMYC1 Leads to Trichome and Root Hair Patterning Defects by Abolishing Its Interaction with Partner Proteins in Arabidopsis. *Journal of Biological Chemistry* 287: 14109-21

Zhao M, Morohashi K, Hatlestad G, Grotewold E, Lloyd A. 2008. The TTG1-bHLH-MYB complex controls trichome cell fate and patterning through direct targeting of regulatory loci. *Development* 135: 1991-99

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF. 2004. Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. *Plant J* 40: 22-34

Zou C, Lehti-Shiu MD, Thibaud-Nissen Fo, Prakash T, Buell CR, Shiu S-H. 2009. Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice *Plant Physiology* 151: 3-15

A1. APPENDIX

Table A1.1: MIQE checklist for authors, reviewers, and editors.[a] (Bustin et al 2009)

2. CHAPTER TWO

Reconstructing and characterizing *Pachycladon* "ancestral" epidermal cell fate pathways with a focus on the trichome initiation pathway

2.1 ABSTRACT

The genus *Pachycladon* (Brassicaceae) originated from the hybridization and polyploidy between two highly diverged lineages during the Pleistocene ~1.6-0.8 mya. Following allopolyploidy, the new lineage differentiated into what is now recognized as ten species with variable ecological niches. Interestingly, the group exhibits a stable karyotypic stasis. This group provides an interesting opportunity to study the fates of gene duplicates in the context of a genetic pathway, a feature of polyploidy that is relatively understudied. The Brassicaceae trichome initiation pathway is a good candidate for this purpose as it is well described from the model plant *Arabidopsis thaliana* and is functionally conserved across the family. This chapter aims to reconstruct the two versions of this pathway that were contributed to *Pachycladon* by its diploid progenitors. In short, I attempted to isolate the two homeologs of the four key trichome initiation genes from each of the four *Pachycladon* species. For each gene of interest, the coding regions of all homeologs obtained were aligned with homologs from other Brassicaceae species, and phylogenetic analyses were used to place the *Pachycladon* copies relative to other members of the family. The results for each gene showed that copies of one homeolog (Type 1) were most closely placed with homologs from *Arabidopsis, Capsella,* and *Camelina*, and copies of the other homeolog (Type 2) were sister to that group. The gene trees were quite consistent in the pattern revealed. The interpretation is that all of the Type 1 homeologs originated from one of the *Pachycladon* progenitors and the Type 2 homeologs originated from the other progenitor. This context allows for future analyses to make molecular comparisons between the progenitors' pathways and investigations into the expression profiles of the different copies of the pathway in *Pachycladon* species.

2.2 INTRODUCTION

2.2.1 Polyploidy and its consequences

Polyploidy or whole genome duplication has long been recognized as a major recurrent evolutionary force in angiosperms (Grant 1981, Soltis et al 2010, Weiss-Schneeweiss et al 2013). Polyploid organisms have more than two complete sets of chromosomes, which results in greater gene content, often increased cell size, and altered transcriptomes and metabolism (Doyle & Coate 2019, Van de Peer et al 2017). Frequently, polyploids confer better resistance to environmental stresses frequently not tolerated by their diploid progenitor species (Bennett 2004, Van de Peer et al 2021).

There is an exhaustive literature featuring the effects of polyploidy on the genetic and genomic constitution of plants (Birchler 2012, Liu et al 2009, Soltis et al 2004, Weiss-Schneeweiss et al 2013). Often genomes are rapidly reorganized soon after polyploidy. At the same time, it is influenced by cytoplasmic–nuclear interactions (Soltis et al 2004, Wendel 2000). Chromosome/chromosomal fragment loss and changes in chromosome organization, gene order, expression, epigenetic modification, and biological network topology altogether lead to genome contraction over time that converts polyploid genomes to functional diploids (Moghe & Shiu 2014, Soltis et al 2004, Wendel 2000). The genetic changes that take place are based on the alteration or loss of genes or DNA sequences. Gene deletion caused by illegitimate recombination or non-homologous recombination is considered the predominant mechanism of fractionation and genome size reduction in flowering plants (Devos et al 2002, Freeling et al 2015, Li et al 2021). The duplicated gene copies can be lost/fractionated or randomly/preferentially retained (Freeling 2009), which is also reflected at the expression level via differential expression or redundant expression and subgenome dominance (Moghe & Shiu 2014). The retained genes can be evolved under adaptive evolution (e.g., neofunctionalization, dosage selection) or purifying selection aiming at maintaining the ancestral functions (e.g., subfunctionalization, dosage balance) (Kondrashov 2012).

Compared to research on the genomic and genetic effects of polyploidy, there is less known about the potential effects on or from genetic pathways. When preferentially retained or

lost genes are related to transcriptional and genomic changes following genome doubling, alterations in genetic complexes, pathways, or networks could occur (Moghe & Shiu 2014). Yet, most cases studied align with the gene balance hypothesis, which explains that the loss of a gene component involved in a complex within a genetic pathway could result in a perturbation of stoichiometric relationships between the other interacting genes in the complex, thereby attempting to establish a new balance among the component genes by mutating the copies of other gene components accordingly (Moghe & Shiu 2014, Shi et al 2015, Van de Peer et al 2017).

To explore gene duplication in the genetic pathway context, we have selected the trichome initiation pathway in the allopolyploid genus *Pachycladon* (Brassicaceae), as described in the previous chapter. While this system affords many advantages and opportunities, it has the challenge of a complicated genome and many duplicated gene families. This chapter describes our approaches taken to reconstruct the 'ancestral' trichome initiation pathways of *Pachycladon*.

2.2.2 *Pachycladon* as a study system

2.2.2.1 Allopolyploid origin and evolution

Pachycladon is an ideal example of an allopolyploid originating from two highly diverged parental lineages from the *Brassicaceae* (Joly et al 2009) (see Chapter 1.3.2 for a detailed explanation). *Pachycladon*'s polyploid origin was first suggested by McBreen and Heenan (2006) reporting the presence of duplicated nuclear genes. Later the allopolyploid state of the genus was reported by Joly et al (2009), analyzing five nuclear genes phylogenetically. All genes were found to be present in two copies, in similar positions in different gene trees compared to the rest of the phylogenies. Thus, it was confirmed the polyploid event was common to different species tested, so it likely occurred before the genus diversified. In addition, the grouping of gene copies with distantly related species in a similar manner among different gene trees supported *Pachycladon*'s allopolyploid origin.

As stated in chapter 1.3.2, parents of the ancestral *Pachycladon* species are not known except for links to evolutionarily close relatives. Joly et al (2009)'s analysis suggested that the

two genomes of *Pachycladon* started to diverge 8.18 ± 4.37 mya. They also estimated that the allopolyploidization event would have taken place at around 1.6 mya and that the parental lineages would have independently evolved for over ~ 6.4 mya before the initial hybridization event. Hence, the parents' genetic distance is thought to be exceptional, showing one of the most extreme cases yet discovered. Based on the research done to date, the phylogenetic positions of the two diverged parental copies were found most likely within lineage I of the *Brassicaceae* phylogeny but distantly related with one copy in the Arabidopsoid clade and the other one within a more basal clade in lineage I likely in the *Smelowskiea* clade (Huang et al 2015, Joly et al 2009, Zhao et al 2010).

The *Pachycladon* lineage is thought to have experienced four whole genome duplication events in its history (Joly et al 2009, Mandáková et al 2017), including three commonly known whole genome duplication events originally described from *A. thaliana* (see chapter 1.1.4) (Blanc et al 2003, De Bodt et al 2005, Vision et al 2000) and *Pachycladon*'s allopolyploidy event. Mandáková et al (2017) found this mesopolyploidy event is shared by *Pachycladon* and *Stenopetalum nutans* (one of the Australian *Microlepidieae* taxa); however, *Stenopetalum* was estimated to have undergone genome evolution at a rate 20% faster than *Pachycladon*. *Pachycladon* is also mentioned as having a slowly evolving genome compared to the Australian *Camelineae* species used in the study reasoning that *Pachycladon*'s perennial growth habit affects its greater chromosomal stasis than the annual Australian *Camelineae*.

2.2.2.2 High rate of species adaptive radiation and karyotype stasis

Hybridization and allopolyploidization together with ecological opportunities have been suggested as stimuli for biotic diversity and adaptive radiation (Schenk 2021). In the case of *Pachycladon*, the high rate of adaptive radiation caused the lineage to spread into distinctive ecological niches found in the Southern Alps (Joly et al 2013, Mitchell & Heenan 2000) and to form a diversity of growth habits as well (see Table 2.1). Notably, the ancestral *Pachycladon* karyotype remained stable during this Pleistocene adaptive radiation of the ancestor into ten different species (Mandáková et al 2010). Species radiation in islands has been studied in depth

and chromosomal stasis or uniformity of chromosome number or ploidy level is a fact that has been exhibited many times (Joshi et al 2022, Price & Wagner 2004, Stuessy & Crawford 1998). Likewise, *Pachycladon* has a relatively slow post-polyploid genome diploidization compared to some other mesopolyploid crucifer taxa studied by Mandáková et al (2017). Four *Pachycladon* species (*P. cheesemanii*, *P. enysii*, *P. exile*, and *P. novae-zelandiae*) have been analyzed using comparative chromosome painting (CCP) (Mandáková et al 2010) and found to have identical karyotype structures. All *Pachycladon* species possess ten chromosome pairs (2n= 20) (Mandáková et al 2010, Yogeeswaran et al 2011). The ancestral *Pachycladon* allopolyploid (n=16), is thought to have originated from two ancestral Crucifer (*Brassicaceae*) karyotypes (ACK, n=8) and then undergone genome reshuffling such as inversions, translocations, and centromere inactivation/ loss, and ended up with n=10 (Mandáková et al 2010). Figure 2.1 below extracted from Mandáková et al (2010) clearly illustrates how the two sets of eight ancestral Crucifer chromosomes brought together in the hybridization event, were arranged within ten chromosomes of the *Pachycladon* ancestor. In other words, 24 conserved genomic blocks belonging to both parents have been distributed among the ten *Pachycladon* ancestral chromosomes as shown in Figure 2.1. However, as *Pachycladon* is a young allopolyploid, whose genome doubling happened around 1.61 mya, the species radiation has been occurring for a relatively short time compared to many paleopolyploids, which explains the few signs of diploidization as suggested by the presence of two potentially functional copies of most nuclear genes (Joly et al 2009).

Table 2.1: Morphological and distributional characteristics of *Pachycladon* species extracted from (Yogeswaran et al 2011)

Species	Growth habit	Leaves	Inflorescence	Siliques	Seed wings	Altitudinal range (mean ± SD) m	Geological parent material of substrate
P. cheesemanii	Polycarpic, woody caudex, subshrub habit	Leaf heterophylly, ovate to broadly elliptic, lobed to serrate; lamina and petiole with branched hairs	Slender and terminal; narrow petals	Terete; seeds uniseriate	Absent	10–1,600 (811 ± 49)	Greywacke, semi-schist, schist, plutonics
P. enysii	Monocarpic, caudex stout and soft	Ovate to ovate-lanceolate, serrate; lamina and petiole with branched hairs	Terminal and very stout; broad petals	Laterally compressed; seeds biseriate	Present	975–2,492 (1,885 ± 42)	Greywacke
P. exile	Polycarpic, caudex woody	Leaf heterophylly, ovate to broadly elliptic, lobed to serrate; lamina and petiole with branched hairs	Slender and terminal; narrow petals	Terete; seeds uniseriate	Absent	25–500 (276 ± 137)	Limestone, calcareous tuff and basaltic breccia, and alluvium
P. fasciarium	Monocarpic, caudex stout and soft	Narrowly linear to narrowly linear-lanceolate; glabrous	Terminal and very stout; broad petals	Laterally compressed; seeds biseriate	Present	884–1,058 (1,021 ± 59)	Limestone
P. fastigiatum	Monocarpic, caudex stout and soft	Narrowly elliptic to lanceolate, serrate; lamina usually glabrous (rarely with simple hairs), or with simple hairs only on the petiole and leaf margin	Terminal and very stout; broad petals	Laterally compressed; seeds biseriate	Present	914–2,031 (1,485 ± 33)	Greywacke, semi-schist
P. latisiliquum	Monocarpic, caudex stout and soft	Narrowly elliptic to lanceolate, serrate; lamina glabrous, usually with simple hairs on the petiole and leaf margin	Terminal and very stout; broad petals	Laterally compressed; seeds biseriate	Present	1,036–1,768 (1,441 ± 22)	Sandstone, marble, limestone, plutonics, argillite, granite
P. novae-zelandiae	Polycarpic, caudex semi-woody	Oblong to broadly elliptic, usually lobed or crenate; lamina and petiole glabrous or hairy, hairs simple or branched	Lateral, more-or-less slender; narrow petals	Laterally compressed; seeds biseriate	Absent	1,080–2,031 (1,587 ± 26)	Semi-schist, schist, plutonics
P. radicatum	Polycarpic, caudex semi-woody	Obovate to broadly elliptic, serrate; glabrous	Lateral, more-or-less slender; broad petals	Laterally compressed; seeds biseriate	Present	No data available	Dolerite and siliceous rock
P. stellatum	Monocarpic, caudex stout and soft	Narrowly elliptic to lanceolate, serrate; lamina and petiole with branched hairs	Terminal and very stout; broad petals	Laterally compressed; seeds biseriate	Present	900–1,371 (1,021 ± 34)	Greywacke
P. wallii	Polycarpic, caudex semi-woody	Oblong to oval, usually lobed or crenate; glabrous	Lateral, more-or-less slender; broad petals	Laterally compressed; seeds biseriate	Present	1,100–1,875 (1,403 ± 37)	Sandstone, semi-schist, schist

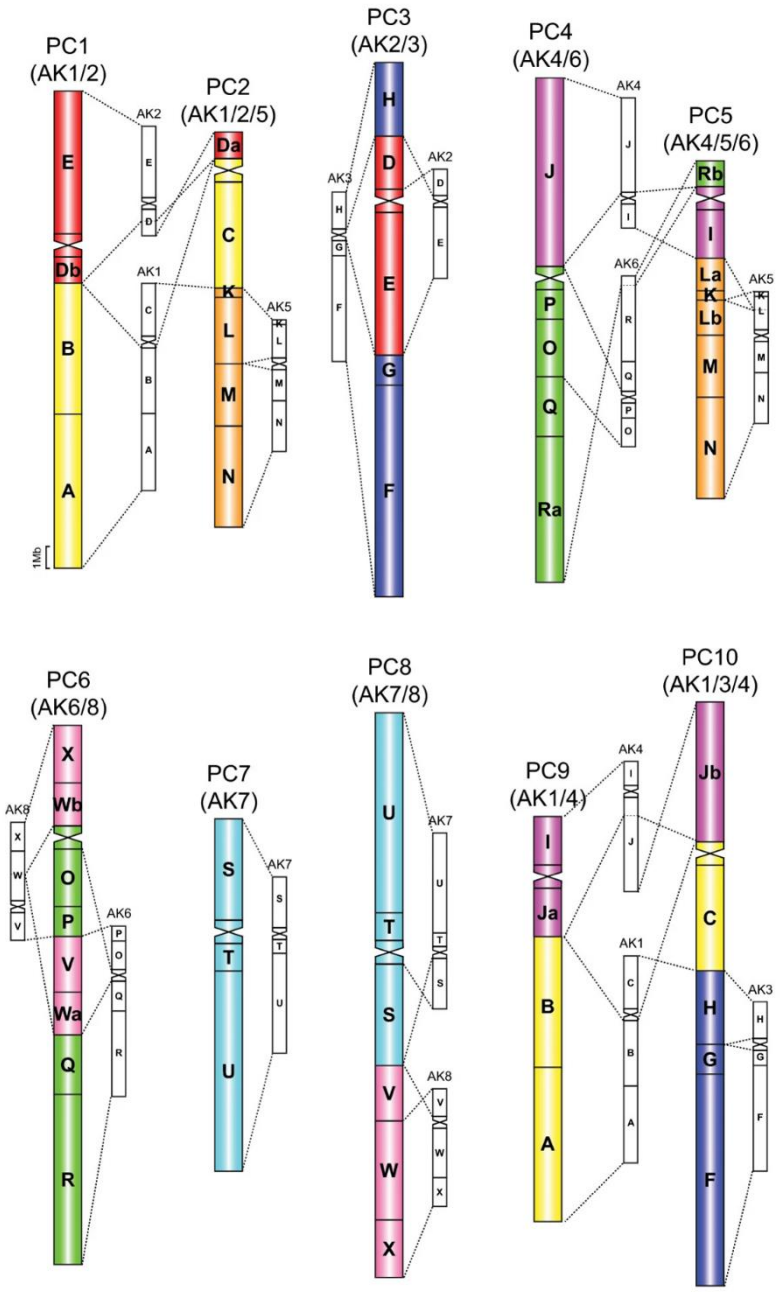


Figure 2.1: Comparative cytogenetic map of duplicated Ancestral Crucifer Karyotype (ACK) (two sets of eight ancestral chromosomes (AK1- AK8)) showing collinearity relationships with ten *Pachycladon* chromosomes (PC1 - PC10). The collinear regions of the two genomes are connected by the dashed lines. 24 conserved genomic blocks (A-X) are found in the ACK which are coloured based on their positions in the chromosomes. Blocks split into two parts are labeled as ‘a’ and ‘b’. The centromeres are shown as sandglass-like symbols, which are colored according to their origin from AK chromosomes (Figure extracted Mandáková et al (2010)).

2.2.2.3 Origin in New Zealand

There are different suggestions on how the ancestral *Pachycladon* lineage dispersed to New Zealand after the hybridization event because it appears unlikely from previous data

(Mitchell & Heenan 2000) that the hybridization event occurred in New Zealand as no *Brassicaceae* in New Zealand was found closely related to *Pachycladon*. A transoceanic dispersal after the hybridization event took place in Eurasia or North America was suggested by Joly et al (2009) because one of the genomes of *Pachycladon* was related to *Transberingia* and *Crucihimalaya* (Al-Shehbaz et al 1999, Price et al 2001). On the other hand, Mandáková et al (2010) suggest an alternative where the parental genomes of *Pachycladon* are closely related, and the allopolyploid event took place in Australia and then dispersed a long distance to New Zealand.

2.2.3 Reconstructing *Pachycladon* "ancestral" trichome initiation pathways

Overall, the origin of *Pachycladon*'s parental lineages is still not well understood although, there is more data to explain they are quite divergent (the paternal parent belongs to *Camelinae* lineage and the maternal parent belongs to *Brassica* lineage) (Huang et al 2015, Joly et al 2009, Zhao et al 2010), other than evidence to say they are originating from the same lineage (Manda´kova´ et al 2010). Therefore, the objectives of this chapter include classifying which copies of each gene go with one another to represent the ancestral pathways and further elaborating on the parental origins within the *Brassicaceae* phylogeny from the view of point of four key genes (*TTG1*, *GL1*, *GL3* and *EGL3*) involved in the trichome initiation pathway (see chapter 1.2.1.2 for detailed description), if good enough sampling from Brassicaceae homologs could be obtained. Here we hypothesize that Type 1 copy of each gene goes with another Type 1 copy of another gene, thus all Type 1 copies originate from one parent, and all Type 2 copies are derived from the other parent. The sequencing work done to obtain full coding sequences of the genes will be mostly discussed here which will lead to molecular evolutionary analyses discussed in the next chapter.

2.3 MATERIALS AND METHODS

2.3.1 Plant material

Pachycladon leaf and seed samples from the populations listed in Table 2.2 below were collected from the South Island of New Zealand. The three *Pachycladon* species, along with *P. cheesemanii* available genome sequence data were selected in the study to cover the phylogenetic range of the genus based on previous studies (Heenan & Mitchell 2003, McBreen & Heenan 2006).

Table 2.2: *Pachycladon* plant materials collected and used.

***Pachycladon* population**	***Pachycladon* species**	**Location collected**	**Accession number**
Pch 002	*P. enysii*	Miromiro Hanmer Springs	#20
Pch 004	*P. novae-zealandiae*	Cardrona skifield	Not reported
Pch 007	*P. fastigiatum*	Mount Hutt	#16

Seeds were sown on autoclaved Daltons Premium Seed Mix soil in pots, watered well, and kept in dark conditions at 4-5 °C for 10 days to overcome seed dormancy, after which, the pots were transferred to a growth room at 20 °C with 16 hours of light and 8 hours of dark and watered approximately once per week. Seedlings grew very slowly and were transplanted after five months into fresh soil topped with Daltons premium-grade pea metals to reduce fungal growth.

2.3.2 Genomic DNA extraction

Young leaves from three-month-old *Pachycladon* plants were used for DNA extractions using a modified CTAB protocol (Doyle & Doyle 1990). From each plant, two to three young leaves (~¾" long) were bashed using a bead mill (Roche MagNA Lyser) at 5,000 RPM for 90 seconds. For some leaves that were difficult to break up (e.g., *P. fastigiatum*), liquid Nitrogen was used, and leaves were ground using a Mortar and pestle. To ground leaf material, 450µL of preheated (65°C) CTAB (2% hexadecyltrimethylammonium bromide), in which 2µL of βME (2-mercaptoethanol) was mixed for each 1mL of CTAB, was added to ground leaf material and vortexed for 5-10 seconds to mix thoroughly. Then the samples were

incubated at 60-65°C for ~30 minutes. In a fume hood, 800µL of Chloroform/Isoamyl alcohol (24:1) was added, mixed vigorously, and spun in a microfuge at 14000 rpm for 5 minutes. The top aqueous layer (300-400µL) was transferred to newly labeled snap-cap tubes. This step was repeated using the aqueous layer. Then again, the top aqueous layer (100-200 µL) was transferred into new labeled tubes. Next, 400µL of ice-cold Isopropanol was added and the tubes were inverted slowly about 5 times. Afterwards, samples were placed in -20°C freezer for ~1 hour. Next, they were centrifuged in the microfuge at 13000 rpm for 10 minutes. After that, the supernatant was slowly discarded without losing the pellet. Then 500µL of room temperature 70% ethanol was added and the tubes were inverted gently ~5 times to rinse the pellets, followed by centrifuging in a microfuge at 14680 rpm (top speed) for 5 minutes. The supernatant was slowly discarded and then the tubes were inverted on Kim-wipes on the benchtop over several hours to overnight to dry. When the pellets were dry, they were resuspended in 100µL of filter-sterilized TE. After the DNA was completely resuspended, the samples were stored in -20°C freezer.

Extracted DNA was visualized in 1% agarose gels in a Bio Rad gel documentation system (UVIDOC HD6 Touch-plus) after staining with 0.5µg/mL Ethidium Bromide. The DNA was also quantified using a Nanodrop (Nanodrop Technologies using the software ND-1000 V3.6.0).

Extracted genomic DNA samples were further processed as described below 1) for genomic DNA Sanger sequencing after amplifying with homeolog-specific primers (2.3.4) and 2) for Illumina miseq sequencing (2.3.5) as summarised in Figure 2.2.

2.3.3 Sequencing methods used to generate gene sequences

Multiple methods were used to get as complete of gene coding sequences as possible. First, following a traditional approach, PCR products were generated, and Sanger sequenced (described in the next section). But, since it was difficult to get complete gene copies, especially with two different homeologous copies for each gene, genomic skim sequencing on an Illumina MiSeq was done for *P. enysii* and *P. novae zealandiae*. For *P. cheesemanii*, there skim sequencing data was already available from previous studies (Dong et al 2019). As leaf

and seed resources were not available for this species at the start of the study, the available skim sequences were used to isolate the genes of interest. Late in the study, with more plant materials available and as part of the validation of primers designed for quantitative expression experiments (described in a later chapter), cDNAs from some samples were Sanger sequenced. In addition, there was an opportunity to do a skim sequencing run of cDNA from *P. fastigiatum*. The sequential process taken to generate sequences accompanying all the methods mentioned here is presented stepwise in the flow diagram (Figure 2.2).

2.3.4 Genomic DNA Sanger sequencing

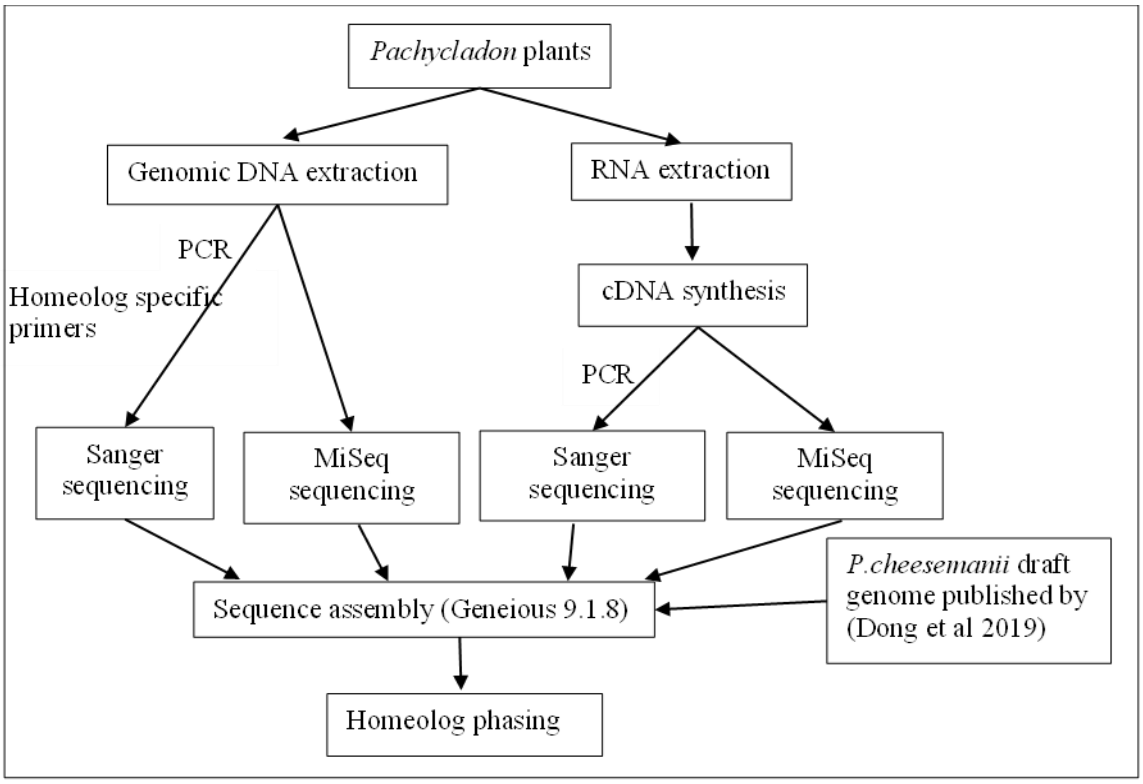


Figure 2.2: Procedure used to generate *Pachycladon* homeolog coding sequences

The first approach to isolating gene sequences used extracted genomic DNA as a template to PCR amplify the genes of interest using homeolog-specific primers. Then the amplified PCR products were used in Sanger sequencing as explained below.

2.3.4.1 Homeolog-specific primer design

The four genes involved in the formation of trichome initiation complex in the pathway- namely *TTG1*, *GL1*, *GL3*, and *EGL3*-were used to design the primers. Initially, conserved regions of the genes based on *Arabidopsis thaliana* and other Brassicaceae sequences and a few *Pachycladon* sequences available from previous studies were used to design primers for the genes of interest. From polymorphic starting sequences or cloned products (described further down), homeolog-specific polymorphisms were used to design homeolog-specific primers to generate full-length coding sequences for each homeolog of each gene of interest from the four study species.

Primer design was mainly done manually, following basic primer design guidelines such as: selecting 18-25 base pairs, including a GC content of around 40%-60%, 2-3 GC bases in the last 5 bases at the 3' end for strong annealing, avoiding four or more mono or di-nucleotide repeats and complementary sequences within and between primer pairs. In some cases, Primer3 (Untergasser et al 2007) was used in Geneious to confirm the primer T_m (melting temperature) is less than the annealing temperature of the PCR program, and hairpin T_m and self-dimer T_m are none or way lower than the annealing temperature of the reaction (Alfandary 2015). Details of the primers used are shown in Table no 2.3.

Table 2.3. Details of used primers; the primer combinations used are as in the order of the list below and sometimes the same forward or reverse primers were used with different reverse or forward primers to make different primer combinations as in the table order.

Gene	Homeolog	Primer name	Primer sequence (5'-3')	Tm (°C)
TTG1	Type 1	Pachy-At5g24520-5pf2-F	AATCAGATCACACATTTGCC	51.0
		Pachy-At5g24520-i4-R	AAATCGTCGCCATGTATCTAAGA	53.9
		Pachy-At5g24520-i2-F	TTGGGATATTGAGAAGTCT	47.4
		Pachy-At5g24520-3pf5-R	ATCACTTCACATCTGCAYC	51.2
		TTG1-02-STOP-R	TCAAACTCTAAGGAGCTGC	51.9
		1367_PCH-TTG1-3pf22-R	AGGTCTATGCTAAATCACTTCA	51.2
		1356_PCH-TTG1-5pf10-F	CGAATCAGATCACACATTTGCC	54.5
		1370_PCH-TTG1-i26-R	ACGGATCCATCAGCAGAGAC	56.6
	Type 2	Pachy-At5g24520-5pf3-F	ATCAGATCAGCCAAYCTARC	51.9
		Pachy-At5g24520-i3-R	GAATCGTAGCCATATATCTYAGG	51.2
		Pachy-At5g24520-i1-F	TTGGGATATCGAGAAGTGC	51.7
		Pachy-At5g24520-3pf1-R	ACGCACCYGACGATTATACRGG	59.0
		1363_PCH-TTG1-5pf14-F	CCGTCAAGAAACCCAATCTCG	56.0
		1362_PCH-TTG1-i13-R	CCCAGAGACGGAGGAAATCG	57.3
GL1	Type 1	Pachy-GL1-i-F1	GAAGACCTCATTATTCGTCTCCAC	55.0
		128-Pachy-GL1-T1-int_F	ATTTATGAGACTTTTTCC	47.6
		124-Pachy-GL1-T2-3fl_R	TAATCTCAGAATTAAGAC	42.7
		1251-PCH-GL1-T2-5fl	CCCTTATACGTATTGAYG	45.4
		1254-PCH-GL1-T12-iR	ACACACMTAGAGRRAGGGTT	54.0
		GL1-136-Pachy-copy2-exon2-R	TCATCCATCTCAGTCTACAA	49.9
		1253-PCH-GL1-T1T2-iF	GAAGACCTCATTATTCGTCTCCAY	54.6
		1469_Pachy_GL1_T1_i_12F	GATTACTTCAATTTTAAAAGGTG	46.4
		1463_Pachy_GL1_T1_i_12F	ATGGCGGAAAATACACAA	49.4
		1256-PCH-GL1-T1-3fl_R	GAATTAAGACAATTCAAA	39.5
		1461_Pachy_GL1_T1_5pF	AGTCACGGAAACCCATTATAAT	51.7
		1468_Pachy_GL1_T1_i_12R	ATTAAAAACTCTTCCACATGACC	51.3
		1459_Pachy_GL1_T1_i_12R	CCACATGACCAATTAACG	48.5
		1465_Pachy_GL1_T1_i_12F	TTATTCGTCTCCACAAGCTCC	54.5
		1471_Pachy_GL1_T1_3pF_R	GACAATTCAAAACAGCAAC	47.8
	Type 2	1250-PCH-GL1-T2-5fl_F	AGTGTGTGCATATAMGCACG	54.2
		1253-PCH-T1T2-F	GAAAAACTGGTACTCTTTCTCTCT	56.3
		1255-PCH-GL1-T2-3fl_R	GAATTAAGACAGTTAAAC	39.9
		1253-PCH-T1T2-F	GAAAAACTGGTACTCTTTCTCTCT	56.3
		125-Pachy-GL2-T2-3fl	ATTAAGACMGTTAAACAC	42.7
GL3	Type 1	Pachy_GL3_F1	CCGTCGCTGAATCTTCTTCC	55.7
		Pachy_GL3_R1	ACGTCGTCGTTTCTTGGATC	55.0
		Pachy_GL3_F2	TCCGTCGCTGAATCCTCAA	56.2
		Pachy_GL3_R2	CTTCATCGTCATCGTCCCAC	55.5
	Type 2	Pachy_GL3_T1_f_1688F	GATCTAGGGATGAAGAAA	47.2
		Pachy_GL3_T1_i_1690R	CCTAGAATATGTTGCTGAT	49.1
		2186_Pch-GL3T2-5'F2	GTTTCATGTAAATGACAAACAC	52.6
		2187-Pch-GL3T2-R1	ATGCTAACCGAATTACCC	51.6
		2326_P. fs_GL3T2 F1	GTGGGATCTAGGGATGAAG	53.7
		2327_P. fs_GL3T2 R1	TACCACTCTGTGTCGGCA	57.8
EGL3	Type 1	Pachy_EGL3_F1	GGAATCTTCTGGTCTGTCTC	52.2
		Pachy_EGL3_R1	CTTGGATCATTCCTGTTGCG	54.1
	Type 2	Pachy_EGL3_F2	GACACCGAGTGGTACTACTTA	53.1
		Pachy_EGL3_R2	CCGGTTTCTTCCTCTTCATC	53.0

2.3.4.2 Polymerase chain reaction (PCR)

Undiluted genomic DNA was used as the template for PCR with 10µM primer working solutions. For most amplifications, the standard Phusion PCR recipe that comes with Phusion High-Fidelity PCR Kit (Thermo Scientific catalog number F630S), was followed to prepare 10 µL individual reaction volumes (Table 2.4).

Table 2.4: Standard Phusion PCR recipe calculated for 10 µL total reaction

Component	**10 µL reaction**	**Final concentration**
5X Phusion buffer (HF)	2.0 µL	1X
10 mM dNTPs (deoxyribonucleotide triphosphate)	0.2 µL	200 µM
10 µM Forward primer	0.5 µL	0.5 µM
10 µM Reverse primer	0.5 µL	0.5 µM
DMSO (dimethyl sulfoxide)	0.3 µL	3%
Phusion DNA polymerase	0.1 µL	1.0 units/ 50 µL PCR
Template DNA	1.0 µL	<250 ng
Sterilized double distilled water (ddH_2O)	5.4 µL	

The PCR program in Table 2.5 was typically used, but PCR profiles were optimized according to different annealing temperatures of the primers used and expected PCR product size. A Biometra T3000 Thermocycler and Maxgene gradient Thermocycler were used for PCR reactions throughout the study. PCR products were observed on 1.5% agarose gels with reference to a 1 kb plus ladder (GeneRuler-Thermo Scientific).

Table 2.5: Thermocycling conditions used for routine PCR

Step		**Temperature**	**Time**
(Lid temperature)		(99°C)	
Initial denaturation		95 °C	3 minutes
30 cycles	Denaturation	95 °C	30 seconds
	Primer annealing	51-59°C	30 seconds
	Primer extension	72°C	0.5-2.0 minutes (depending on amplicon length and set approximately 1 minute for 1000 bps)
Final extension		72°C	10 minutes
Hold		4°C	

When multiple PCR products were observed, cloning and/or band isolation through gel purification was/were used to screen out the expected PCR product. Such multiple PCR products were cloned into a TOPO vector plasmid (pCR 4- TOPO) using *E. coli* grown in

40mg/L kanamycin LB agar medium following the 'Invitrogen TOPO TA Cloning Kit for Sequencing' protocol. Since Phusion polymerase amplified PCR products were used for cloning, another non-proofreading polymerase (NEB taq polymerase) was used to add 3' overhanging 'A's (Table 2.6) to facilitate ligation to the plasmid vectors by incubating at 72°C for 20 minutes to be used in the cloning reaction within one hour. The transformation of the plasmids into 'One Shot TOP10' chemically competent *E. coli* cells was performed following the one-shot chemical transformation protocol. Following transformation, cells were plated on selective media and incubated overnight at 37°C. Next, colony PCR was done to screen for colonies transformed with the desired insert. 20 µL Phusion PCR reactions were prepared, adding half of the colony as the template. Mostly the same gene-specific PCR primer pairs were used but occasionally, M13 forward/M13 reverse or T3/T7 universal primer pairs were used. The same PCR program, which gave the multiple PCR product, was used for the colony PCR.

Table 2.6: 'A' tailing with *Taq* polymerase

Component	**20.0 µL reaction**
Sterilized double-distilled water (ddH_2O)	11.5 µL
10X Thermopol buffer	2.0 µL
10 mM dNTPs	0.5 µL
Phusion PCR product	5.0 µL
NEB taq polymerase (5U/µL)	1.0 µL

In addition to *E. coli* cloning, gel extraction was used to separate the desired PCR product (when it was of good band intensity) from some multiple products, using Zymoclean Gel DNA Recovery Kit from Zymo Research. 10 µL of the PCR products were loaded in 1.5% agarose gels, stained with 0.5 µg/mL Ethidium Bromide and then the required PCR band was extracted while visualizing the gel under UV light on a Transilluminator (Invitrogen Safe Imager). After that, the gel DNA recovery was carried out using the kit manufacturer's instructions and DNA was eluted in 15 µL of elution buffer. The recovered DNA was run again in an agarose gel to confirm the extraction and isolation of the desired single product, and in the

cases where DNA concentration was low, the PCR was repeated using the gel-extracted product as template, under the same conditions.

2.3.4.3 Sanger sequencing of the homeologs from genomic PCR products from *TTG1, GL1, GL3* and *EGL3* genes from *P. enysii, P. fastigiatum* and *P. novae-zealandiae*

Single PCR products were purified using the QIAquick PCR purification kit (Qiagen, Inc.) following the kit protocol. Purified DNA concentration and purity were estimated on a Nanodrop spectrophotometer, and values in the range of 7.2ng/µL-11.10ng/µL for concentration and 1.78-2.27 for 260/280 ratios were recorded.

The sequencing protocol used in the Doyle lab (modified by Fronny Plume for use in the LoST lab) was used. The sequencing reaction mix was prepared on ice, minimizing exposure to light, adding 2.8 µL of 5X ABI sequencing buffer, 0.8 µL of ABI BigDye, 1.0 µL of 5 µM primer (forward or reverse primer used in the PCR), cleaned PCR template (volume based on product size and concentration; 2.5 ng of DNA was used for every 100 bps of PCR product), and sterilized double distilled water to a total volume of 12 µL. Then the reaction mix samples were processed in the cycle sequencing reaction in a thermocycler following an initial denaturation at 96 °C for 4 minutes, then 25 cycles of denaturation at 96 °C for 10 seconds, primer annealing at 57 °C for 5 seconds, and primer extension at 60 °C for 3 minutes, followed by a hold at 10 °C. 10 µL of the sequenced samples were submitted to Massey Genome Service (Palmerston North), for clean-up reactions (using ABI X-Terminator clean-up method) and then capillary seperation in an ABI3730XL DNA Analyzer.

2.3.5 Illumina MiSeq sequencing of total genomic DNA belonging to *P. enysii, P. fastigiatum* and *P. novae-zealandiae*

Although Sanger sequencing enabled the retrieval of complete sequences for some homeologs, it was difficult to fully sequence some of the other homeologs, due to various

reasons including DNA quality, contaminations, primer issues, PCR product purification, etc. Therefore, genomic skim sequencing using the Illumina MiSeq platform was initiated as an alternative to fill the sequencing gaps.

Extracted genomic DNA (explained in 2.3.2) was run in a 1% agarose gel with 1kb plus standard ladder to confirm its integrity. DNAs also were submitted to the MGS for genomic DNA LabChip assay to identify the best DNAs for library preparation and sequencing. quality and quantity assessment.

For one sample of each species, library preparation was performed using the Illumina Nextera Flex DNA Library Preparation Kit_V1. Each sample was indexed separately so that they could be pooled with other samples and their sequences could bioinformatically separate following the MiSeq run. The library insert sizes obtained were in the range of 433-588 bp. The Illumina MiSeq 500 cycle Kit_V2 (2x 250 base paired-end) was used as the sequencing technology; this usually generates ~6-7.5 Gb per run with ~24-30 million paired-end reads (Illumina prepared PhiX control library was used as the control). The library quality was checked using the LabChip DNA high-sensitivity assay, and PerkinElmer Victor Plate Reader.

The data output obtained from the MGS was in the form of raw sequences, processed sequences, processed-trimmed, and QC data, all in fastq format. Raw sequences were subjected to adapter removal using the fastq-mcf program from the ea-utils suite of tools to obtain processed sequences. Then, processed reads were trimmed to their longest contiguous segment based on quality scores less than a quality cut-off, set at 0.01, using the dynamic trim application from SolexaQA++ software. Finally, for the QC analysis, the original sequences were run in the processes, SolexaQA++, fastQC, and fastQscreen. This initial processing was carried out by the MGS staff.

2.3.6 RNA extraction, cDNA synthesis, cDNA Sanger sequencing, and cDNA MiSeq sequencing

Although genomic DNA Sanger sequencing and MiSeq sequencing enabled retrieving complete coding sequences of *TTG1* and *GL1* homeologs, those methods did not yield complete sequences of *GL3* and *EGL3* coding regions, due to the long length of those genes,

unavailability of discrete homeolog-specific primers, and high sequence similarity among the paralogs. cDNA sequencing (Sanger/MiSeq) was identified as a more feasible method for these rather complex genes. First, RNA extraction and cDNA synthesis were done as described in the Materials and methods of Chapter 4 (4.3.1), to process for sequencing.

Similar to genomic DNA, cDNA was used as the template for PCR amplification with available homeolog-specific primers, cleaned up, and used in Sanger sequencing. An opportunity was available for a skim cDNA sequencing run on a single sample and we selected *P. fastigiatum*, as, at the time, this species had the most missing homeolog sequences. For this, double-stranded (ds) cDNA was required to align with other external samples used in the same run. Single-stranded cDNA generated using Invitrogen SuperScript III First-Strand Synthesis System for RT-PCR (chapter 4.3.1) was converted into ds cDNA using NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module (Catalog # E6111S) as per the instructions in the kit up to the step 2.9. After that, cDNA MiSeq was carried out similarly to the genomic MiSeq run described above.

Overall, all the sequences obtained from multiple methods described above (figure 2.2) were used hereafter for sequence assemblies and further processing as described below.

2.3.7 Sequence analysis

2.3.7.1 Assembly of reads

Processed and trimmed MiSeq data were analyzed using Geneious 9.1.8 (Geneious 9.1.8 2017). First, processed reads were paired using the ‘set paired reads’ tool (based on R1 and R2 universal primers which constructed the paired-end reads). Then, those paired reads were merged using ‘merge paired reads’; this results in two files: merged processed and unmerged processed (the latte contains paired sequences that could not be merged due to nonoverlapping sequences). These merged and unmerged reads were then assembled via de novo assembly utilizing Velvet short read assembler (Zerbino & Birney 2008), run as a plug-in in Geneious. Velvet optimizer was run selecting the default k-mer length range from 27-35 (which uses k-mer lengths from 27 to 35, 8 times to find out the length giving optimal

assembly). As a result, the assembly generates several consensus reads, generally ~200,000 bp in length.

In addition, Sanger sequences were also trimmed to remove poor quality sequences including high background noise and multiple signals, and then the filtered sequences were trimmed at the ends to retain only good quality sequences to process for alignment.

2.3.7.2 Mapping MiSeq reads to the *Arabidopsis thaliana* homolog

Arabidopsis thaliana was identified as the best species to use as a reference for *Pachycladon* genes due to its close phylogenetic relationship with one of the *Pachycladon* ancestors (Joly et al 2009). Filtered *Pachycladon* sequence sets obtained from all methods described above (summarised in Figure 2.2) for each species were separately mapped to the *A. thaliana* homologs of each gene of interest; The *A. thaliana* sequences were retrieved from TAIR (The Arabidopsis Information Resource) (Huala et al 2001). The Geneious mapper was used to map the sequences to the reference, using iterative fine-tuning (up to five times) (Geneious calls the consensus sequence of the reads and repeats the mapping process taking the consensus sequence as the reference recurrently).

2.3.7.3 Phasing homeologs

Due to the allopolyploid nature of *Pachycladon*, two copies were expected for each gene of interest. Therefore, each collection of sequences that mapped to a particular gene was further phased into two groups, Type 1 and Type 2, based on polymorphisms in sequence overlaps. Preliminary results revealed that one homeolog is always much more similar to the *A. thaliana* homolog than the other and homeologs in *Pachycladon* tend to be ~9% different in sequence identity.

MiSeq merged and unmerged sequence data were aligned to *A. thaliana* reference homologs. First, the reads were mapped to *A. thaliana* customizing the Geneious mapper sensitivity to set the minimum overlap identity to 85%, allowing mapping of both types of *Pachycladon* reads. Then, considering the polymorphisms in the sequence overlap, the reads

were phased as Type 1 (more similar polymorphisms to *A. thaliana*) and Type 2. The reads belonging to two types were extracted separately and consensus sequences were extracted. Then all the reads were again mapped to *A. thaliana* setting the minimum overlap identity to 95%, expecting the resulting alignment to contain only Type 1 reads, to validate the Type 1 consensus sequence extracted above. The percentage similarity thresholds were selected based on preliminary manual homeolog phasing done on Sanger sequence alignments using type specific polymorphisms (more similar and more dissimilar polymorphisms compared to the reference *A. thaliana*) and then separating sequences into two types.

After generating consensus sequences for each gene homeolog using the sequence assemblies obtained using different sequencing methods (e.g. for *P. enysii*, *TTG1* Type 1, separate assemblies were generated using genomic Sanger sequences and genomic MiSeq sequences), they were combined to confirm that different methods yielded identical homeologous consensus sequences. When allelic variations were observed, they were traced back to original sequence assemblies to confirm if it was a true allelic variation or a sequencing artifact. Individual homeolog assemblies were also manually confirmed. At last, species Type 1 and Type 2 consensus sequences were extracted, and the coding sequences were generated only based on alignment with annotated *A. thaliana* sequences. All the sequencing methods used to generate and validate gene homeologs of all species are shown in Table 2.7. While complete assemblies were made for Type 1 and Type 2 homeologs for all species for *TTG1* and *GL1*, the longer bHLH encoding genes, *GL3* and *EGL3*, proved more difficult and some gaps remained.

Table 2.7: Summary of all the methods utilized to generate and validate *Pachycladon* final consensus gene homeologous sequences. i.e., *Pachycladon* species names are abbreviated (*enysii* = *en*, *fastigiatum* = *fs*, *novae-zealandiae* = *nz* and *cheesemanii* = *ch*), Type 1 is denoted by T1 and Type 2 by T2. **P.ch* published data refers to (Dong et al 2019).

Gene	***Pachycladon* species and homeolog copy**	**Sequencing methods used to generate the final consensus**			
		Sanger sequencing		Illumina miseq	
		Genomic	cDNA	Genomic	cDNA
TTG1	*P. en* T1	√	√ (qPCR product seq)	√	
	P. en T2	√		√	
	P. fs T1	√	√ (qPCR product seq)	√	√
	P. fs T2	√	√ (qPCR product seq)	√	√
	P nz T1	√		√	
	P nz T2	√		√	
	P. ch T1			√ (from *published data)	
	P. ch T2			√ (from published data)	
GL1	*P. en* T1	√	√ (cDNA seq and qPCR product seq)	√	
	P. en T2	√	√ (cDNA seq and qPCR product seq)	√	
	P. fs T1	√		√	
	P. fs T2	√		√	
	P nz T1	√		√	
	P nz T2	√		√	
	P. ch T1			√ (from published data)	
	P. ch T2			√ (from published data)	
GL3	*P. en* T1	√	√ (cDNA seq and qPCR product seq)	√	
	P. en T2	√	√ (qPCR product seq)	√	
	P. fs T1	√	√ (cDNA seq)	√	√
	P. fs T2	√	√ (qPCR product seq)	√	√
	P nz T1	√	√ (qPCR product seq)	√	
	P nz T2	√	√ (qPCR product seq)	√	
	P. ch T1			√ (from published data)	
	P. ch T2		√ (qPCR product seq)	√ (from published data)	
EGL3	*P. en* T1	√	√ (qPCR product seq)	√	
	P. en T2	√	√ (qPCR product seq)	√	
	P. fs T1			√	√
	P. fs T2	√		√	√
	P nz T1			√	
	P nz T2	√		√	

	P. ch T1		√ (qPCR product seq)	√ (from published data)	
	P. ch T2		√ (qPCR product seq)	√ (from published data)	

2.3.8 Phylogenetic analysis

One of the key objectives of this study was to identify which homeologs of each gene go together to represent the 'progenitor' pathways. To this end, we took a phylogenetic approach. Even though the progenitors of *Pachycladon* are unknown (and likely extinct), the initial assessment of divergence between Type 1 and Type 2 homeologs indicated that the assembly of individual gene trees that include both homeologs and homologs from a wide selection of Brassicaceae species would allow relative placements. If the placement of Type 1 and Type 2 homeologs was consistent across gene trees, this would allow for the identification of homeolog sets that correspond with the two progenitors.

First, the *Pachycladon* homeologs were used as input for a BLAST (Basic Local Alignment Search Tool) search (Altschul et al 1990) from the National Center for Biotechnology Information (NCBI), to find homologs consisting of complete coding sequences from the Brassicaceae, and from the outgroup family Cleomaceae The selected BLAST algorithm was 'Megablast for highly similar sequences'.

For each gene of interest, the two coding region homeologs from each *Pachycladon* species (the data sets were less complete for *GL3 and EGL3*), the Brassicaceae homologs, and the Cleomaceae (*Tarenaya hassleriana*) were aligned using the MUSCLE alignment algorithm. For *GL3* and *EGL3* genes, only *Pachycladon* homeologs with complete coding sequences were included in the alignments. Then, phylogenetic trees were constructed by using the Maximum Likelihood method and General Time Reversible model in MEGA 11.0.13 (Tamura et al 2021). The phylogeny was tested using 1000 bootstrap replications. Gamma distribution was selected as the rate among sites and '4' was selected as the discrete gamma category. Subtree-pruning-regrafting-fast (SPR level 3) was selected as the maximum likelihood heuristic method and the initial tree for maximum likelihood was automatically generated by the Neighbor Joining method. A weak branch swap filter was chosen to set more

exhaustive optimization concerning branch lengths and improvements in log-likelihood values. Each tree was rooted using the *Tarenaya hassleriana* (Cleomaceae) homeolog, as the closest outgroup species to the Brassicaceae (Nikolov et al 2019).

2.4 RESULTS

2.4.1 Homeolog sequencing of *TTG1*, *GL1*, *GL3*, and *EGL3* from *P. enysii*, *P. fastigiatum* and *P. novae-zealandiae*

2.4.1.1 Sanger sequencing

Many homeolog-specific gene sequences for all three species were obtained from the traditional PCR/Sanger sequencing approach. However, as this was a very slow and cumbersome approach, especially with the long and complex *bHLH* genes, we also employed a genomic skim sequencing approach. In fact, for most homeologs, Sanger sequencing produced most of the length of complete coding sequences, and skim sequencing was used to fill the gaps found in Sanger sequencing and validate the existing sequences.

2.4.1.2 Illumina MiSeq sequencing

DNA concentrations of the three samples used for the MiSeq runs and the average library insert sizes obtained are shown in Table 2.8.

Table 2.8: DNA concentrations of *Pachycladon* species used for Illumina MiSeq and average library sizes obtained

Sample name	**Qubit DNA concentration (µg/mL)**	**Average library insert size (bp)**	**Number of reads generated**	
			merged processed reads	unmerged processed reads
P. enysii	6.66	433	4,322,624	3,211,316
P. fastigiatum	11.30	462	3,280,394	3,613,662
P. novae-zealandiae	2.80	588	796,279	8,772,274

MiSeq 'processed trimmed data' reads obtained for *P. enysii*, *P. fastigiatum* and *P. novae-zealandiae*, were processed in Geneious 9.1.8 to obtain merged processed and unmerged processed reads. Meanwhile, *P. cheesemanii* sequencing reads were extracted from already published data (Dong et al 2019) generated from 483,792,966 starting raw reads.

2.4.1.3 Sequence analyses

The sequence variation between homeologous Types of the same species was greater than the variation within the same Type of homeologs among different species for all genes

examined (see Table 2.9). The percent identity values of the sequence pairs were extracted from Geneious 9.1.8 summary statistics. Based on the values, the percent identity between the same type of gene copies within *Pachycladon* species or even between *Pachycladon* species is more than 95.2%. In contrast, the similarity between different types within species or between species is in the range of 88.6-90.7%.

Table 2.9: Percent identity values between homeologous sequences. Grey-colored boxes show percentage identity values between Type 1 and Type 2 homeologous sequences. Some values were not calculated due to the lack of complete sequences as indicated. Identities relative to the *A. thaliana* ortholog at shown in the last column.

Gene	***Pachycladon* homeologs (Type 1 or Type 2)**	**Percent identity values between *Pachycladon* homeologs (%)**							**Reference**
		P.en 1	P.fs 1	P.nz 1	P.ch 1	P.en 2	P.fs 2	P.nz 2	*A. thaliana*
TTG1	P.en 1								92.7
	P.fs 1	100							92.7
	P.nz 1	99.22	99.23						92.6
	P.ch 1	99.2	99.2	99.2					92.8
	P.en 2	90.63	90.5	90.5	90.4				89.1
	P.fs 2	90.3	90.43	90.3	90.2	99.81			88.9
	P.nz 2	90.3	90.3	90.18	90.2	99	98.8		88.8
	P.ch 2	90.4	90.4	90.2	90.3	98.9	98.7	98.6	88.9
GL1	P.en 1								91.2
	P.fs 1	99.86							91.4
	P.nz 1	98.56	98.7						91.8
	P.ch 1	98.4	98.6	99					91.8
	P.en 2	88.6	88.7	89.4	88.9				88.3
	P.fs 2	87.3	91	88.1	87.7	98.82			87.2
	P.nz 2	90.3	90.4	91.25	90.7	99.25	98.1		89.8
	P.ch 2	90.6	90.7	91.1	91.2	96.5	95.2	98.7	89.9
GL3	P.en 1								94.4
	P.fs 1	99.9							94.3
	P.nz 1	98.8	98.7						94.1
	P.ch 1	99	98.9	98.7					94.1
	P.en 2	The sequence is not complete							
	P.fs 2	The sequence is not complete							
	P.nz 2	The sequence is not complete							
	P.ch 2	91.3	91.2	91	91.1				89.7
EGL3	P.en 1	The sequence is not complete							
	P.fs 1	The sequence is not complete							
	P.nz 1	The sequence is not complete							
	P.ch 1								93.1
	P.en 2	The sequence is not complete							
	P.fs 2	The sequence is not complete							
	P.nz 2	The sequence is not complete							
	P.ch 2				91.1				90.8

While both homeologs of each gene were isolated from all four species, extracting the two homeologous copies of *GL3* and *EGL3* was more difficult as these are complex genes ~3000-3500 bp with long introns and exons. Importantly, *GL3* and *EGL3* belong to the basic

helix-loop-helix (bHLH) gene family. These two genes have high sequence similarity and are closely related to two other paralogs, *TT8* (Zhang et al 2003) and *AtMYC1* (Symonds et al 2011). These are considered paralogs with highly overlapping functionalities (Bernhardt et al 2003, Zhang et al 2003). With four genes of some similar sequences duplicated in *Pachycladon*, mapping of short sequence reads proved difficult. Therefore, as an approach to distinguish and extract short reads only belonging to *GL3* and *EGL3*, they were aligned with all four bHLH genes to remove reads belonging to other bHLH genes, using unique SNPs belonging to reference *GL3* and *EGL3* genes. Also, some available preliminary *GL3* and *EGL3* short sequences from Sanger sequencing (not assigned to any homoeologous Type though) were used as references to validate between genes. Moreover, *GL3* and *EGL3* introns exhibit extreme variation relative to the reference and between Types, and even within Types, it seems arduous to link between same type exons to build up the full gene copy belonging to a Type. This was overcome after sequencing full-length *P. enysii* Type 1 cDNA (from RNA) and using it as a reference to validate the linkage between exons of Type 1 and therein by confirming the other Type by polymorphisms different from Type 1.

Owing to the complexity of building up the sequences of *GL3*, there were still some sequencing gaps in *GL3* Type 2 *P. enysii*, *P. fastigiatum*, and *P. novae-zealandiae*, whereas complete coding regions were sequenced for *P. cheesemanii* Type 2 sequence and all the Type 1 sequences. Only one-third of the complete coding sequence was able to sequence for *P. enysii* and *P. fastigiatum*. In the *P. nova-zealandiae* Type 2 sequence, the first 36 base pairs were missing at the start of exon 2. When the *GL3* alignment was translated into amino acid sequences, the Type 2 *P. novae-zealandiae* and *P. cheesemanii* proteins seem to have accumulated a large number of mutations including the start codon in *P. cheesemanii*. Premature stop codons were identified in *P. enysii* and *P. fastigiatum* sequences early in exon 2 and *P. novae-zealandiae* and *P. cheesemanii* in exon 5. All the species Type 1 sequences were rather conserved with very few polymorphisms. Thus, it seems the Type 2 copy of all the species is in the process of pseudogenization. *EGL3* sequences were also exhibiting similar

complexities and as a result, complete coding region sequences from only *P. cheesemanii* were obtained. Variable-sized sequence gaps from 100-500 bp were found in all other homeologs.

2.4.2 Phylogenetic analysis of the genes to reconstruct progenitor pathways

A set of 23 *TTG1* sequences that represent the two complete coding region homeologs from each of the four *Pachycladon* species, 14 *Brassicaceae* homologs, and one sequence from the outgroup, *Tarenaya hassleriana* (Cleomaceae), were phylogenetically analyzed to infer the placement of the *Pachycladon* homeologs. The best-scoring maximum likelihood tree generated by MEGA 11.0.13 is presented in Figure 2.3 below.

The *Pachycladon TTG1* gene tree demonstrates gene sequences that belong to at least four *Brassicaceae* lineages out of five shown in the recent *Brassicaceae* phylogeny (Nikolov et al 2019), with three of them having good bootstrap support (100% for the common ancestor of lineage I and lineage IV, and the same for lineage II, but just 62% for lineage I). Due to the more limited sampling of the family (based on the criterion of using only complete CDS sequences) and the examination of a single gene, the overall order of lineage positioning revealed here, understandably, does not exactly match with the published family phylogeny (Nikolov et al 2019).

Pachycladon TTG1 Type 1 homoeologs clustered together are most closely related to sequences from *Capsella*, *Camelina*, and *Arabidopsis* species. This finding is consistent with the inferred *Brassicaceae* lineage I of Beilstein et al (2006) and Nikolov et al (2019) which included *Capsella*, *Camelina*, and *Arabidopsis* species. *Pachycladon* Type 2 sequences form a monophyletic group sister to lineages I and IV. Considering the branch lengths of the tree, Type 1 has evolved more compared to Type 2.

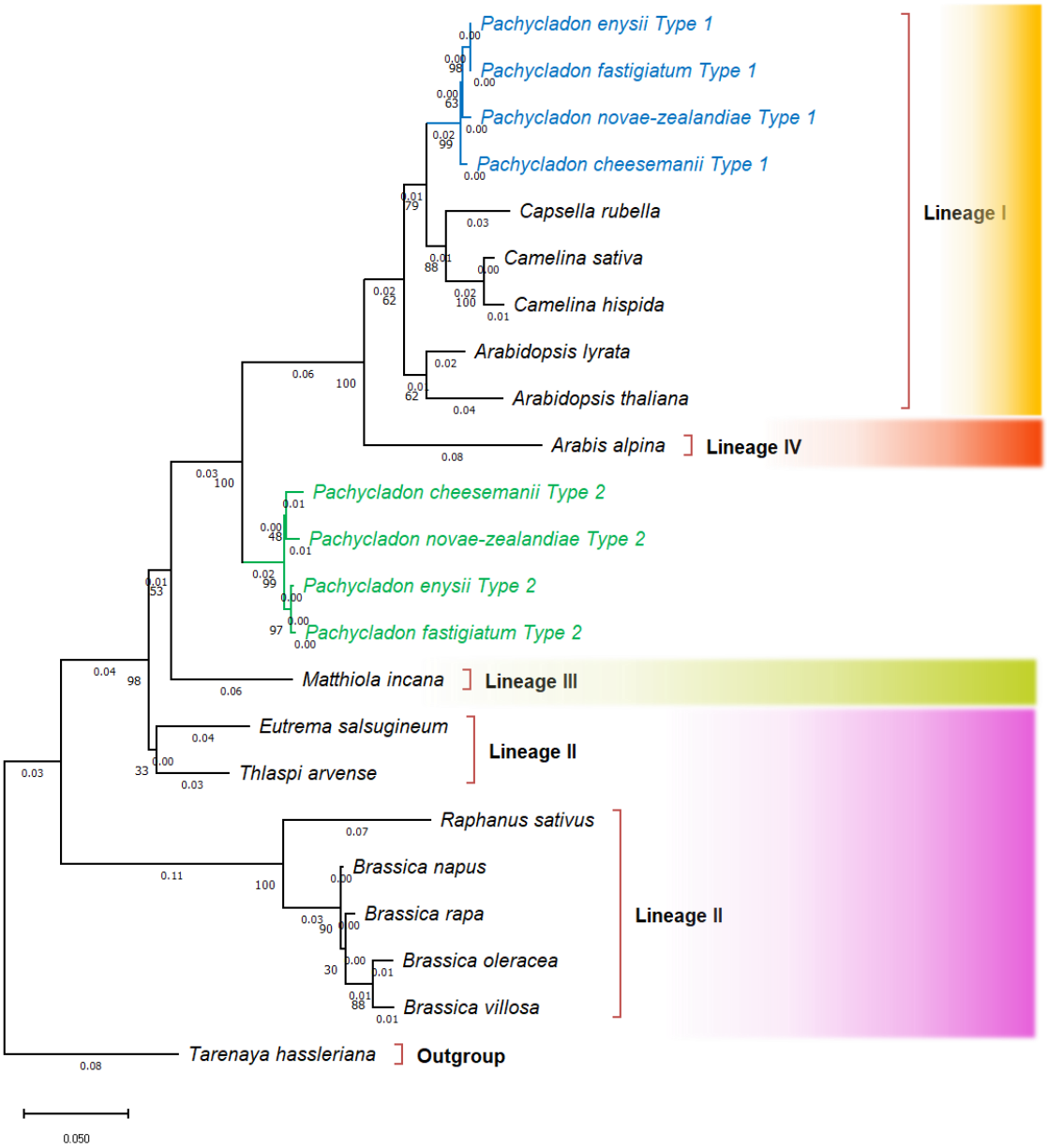


Figure 2.3: The *TTG1* evolutionary history inferred by using the Maximum Likelihood method and General Time Reversible model (Nei & Kumar 2000). The tree with the highest log likelihood (-5470.85) is shown here. The percentage of trees in which the associated taxa clustered together is shown below the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 0.3648)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (below the branches). This analysis involved 23 nucleotide sequences. There was a total of 1041 positions in the final dataset. Evolutionary analysis was conducted and, the phylogenetic tree and the caption were generated in MEGA11 (Tamura et al 2021).

BLAST results for *GL1* resulted in a similar data set of 21 complete CDS sequences; the main difference is that no Lineage III representative was found. Although the *P. fastigiatum*

GL1 type 2 copy produces a truncated protein, the complete sequence comparable to other coding sequences of homologous sequences was used in the alignment to construct the phylogenetic tree.

The *GL1* phylogenetic tree contains only three lineages based on Nikolov et al (2019)'s family-wide phylogeny, but again lineage relationships are not the same. Interestingly, the one lineage IV sequence, represented by *Arabis alpina*, is positioned differently in the *GL1* tree, compared to the *TTG1* tree.

The phylogenetic placement of *Pachycladon GL1* Type 1 sequences was consistent with the results from the *TTG1* phylogenetic tree, clustering with sequences from *Arabidopsis*, *Camelina*, and *Capsella* species. Also similar to the *TTG1* results, the *Pachycladon* Type 2 sequences formed a monophyletic group that is sister to the Type 1 and Lineage I groups (noting the different placement of the *Arabis alpina* sequence). This result is similar to the previously inferred positions of *Pachycladon* homeologs in other studies (Huang et al 2015, Joly et al 2009, Zhao et al 2010) with good support of 99%.

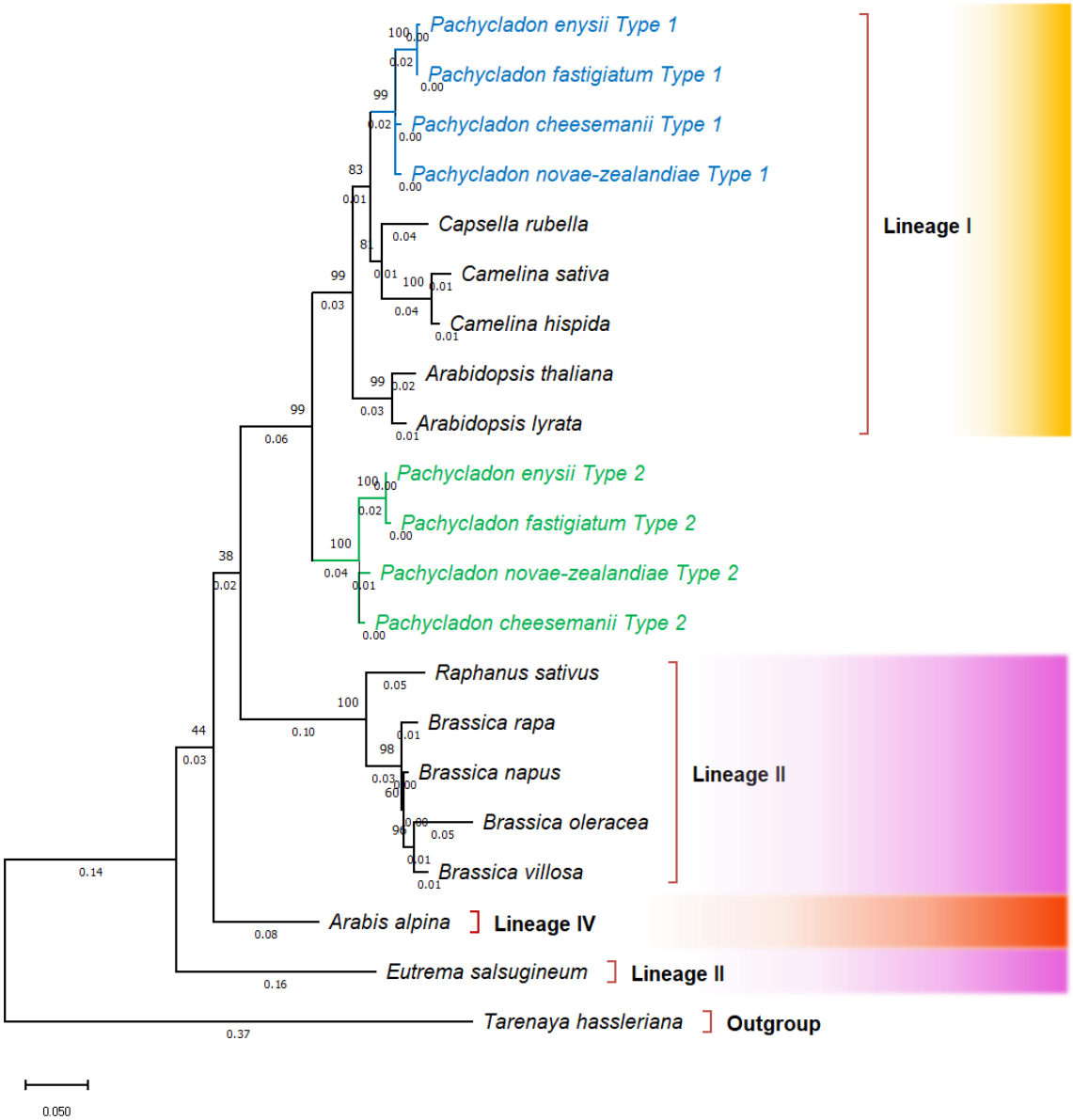


Figure 2.4: *GL1* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model (Nei & Kumar 2000). The tree with the highest log likelihood (-4437.39) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 0.9007)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (below the branches). This analysis involved 21 nucleotide sequences. There was a total of 729 positions in the final dataset. Evolutionary analysis was conducted and, phylogenetic tree and the caption were generated in MEGA11 (Tamura et al 2021).

The relative placements of the *GL3* Type 1 sequences and the lone Type 2 *P. cheesemanii* sequence are again similar to the gene trees described above, with the Type 1 sequences clustering together and with the Brassicaceae Lineage I species and the Type 2 *P.*

cheesemanii sequence sister to that group. Here again, the *Arabis alpina* sequence from lineage IV occupies a position that is different again from the positions in the *TTG1* and *GL1* trees.

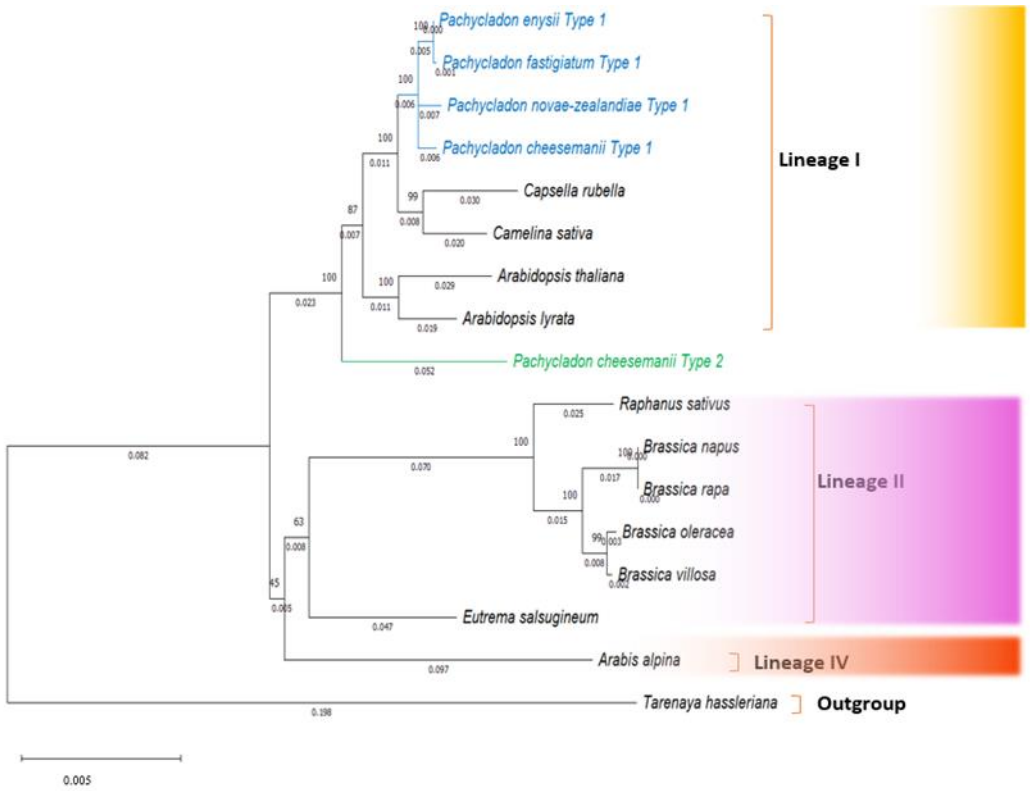


Figure 2.5: *GL3* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model (Nei & Kumar 2000). The tree with the highest log likelihood (-8532.63) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 0.7174)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (below the branches). This analysis involved 17 nucleotide sequences. There were a total of 2012 positions in the final dataset. Evolutionary analysis was conducted and, the phylogenetic tree and the caption were generated in MEGA11 (Tamura et al 2021).

For *EGL3*, complete CDS sequences could only be generated from the larger sequence set available for *P. cheesemanii*. While the lone Type 1 *Pachycladon* sequence was again found to be most similar to Brassicaceae Lineage I sequences, the lone Type 2 *Pachycladon* sequence was not found to be sister to that group as the Type 2s had been for the other three genes. Lineage III *Matthiola incana* is the most basal group of the tree after the outgroup which is

similar to recent Brassicaceae phylogeny (Nikolov et al 2019).

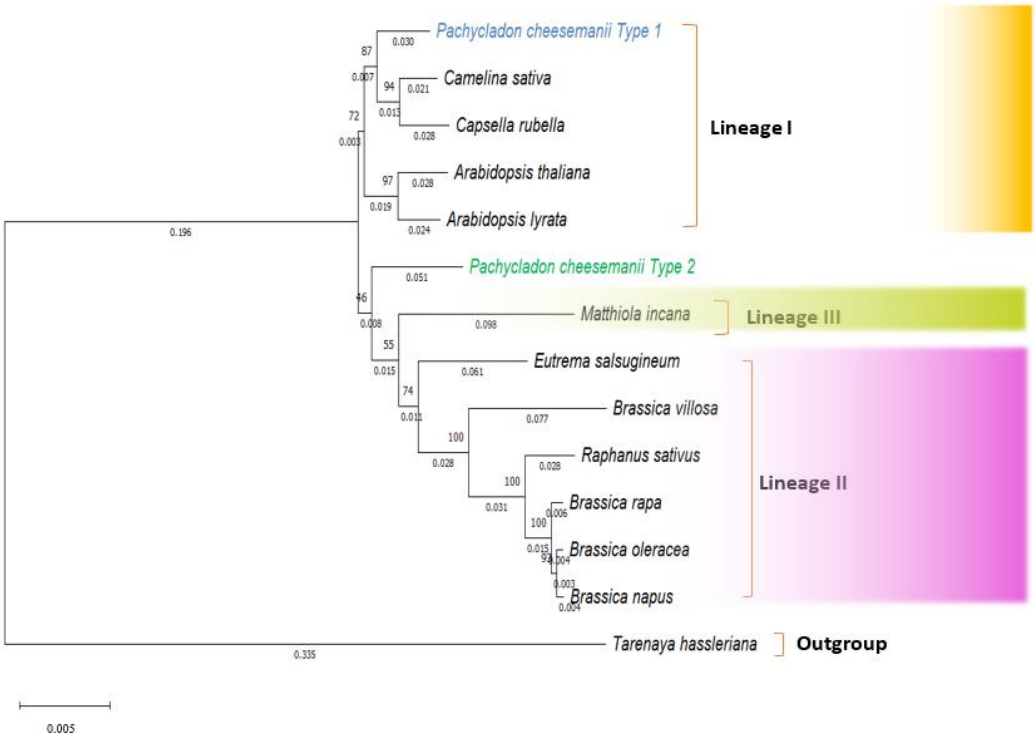


Figure 2.6: *EGL3* evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model (Nei & Kumar 2000). The tree with the highest log likelihood (-9544.74) is shown. The percentage of trees in which the associated taxa clustered together is shown below the branches. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+*G*, parameter = 0.5495)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (below the branches). This analysis involved 14 nucleotide sequences. There were a total of 2043 positions in the final dataset. Evolutionary analysis was conducted and, phylogenetic tree and the caption were generated in MEGA11 (Tamura et al 2021).

Here, it is clear that the phylogenetic placements of Type 1 and Type 2 clusters in all four gene trees, namely *TTG1*, *GL1*, *GL3*, and *EGL3* are consistent. Overall, the common features in the gene phylogenies are that Type 1 homeologs of all species are grouped together with *Capsella*, *Camelina*, and *Arabidopsis* species in lineage I, and Type 2 homeologs of all species are grouped outside the above-mentioned lineage I cluster being less diverged. We could not find any close relatives of Type 2 group within the resulting homologs in our Brassicaceae nucleotide blasts. However, the phylogenetic trees were useful in tracing back the

ancestral origins of the type-specific homeologs which is vital to proceed with their molecular evolution and expression analyses.

2.5 DISCUSSION

Following polyploidy, the fate of gene duplicates, particularly in a genetic pathway context, remains underexplored. Here, we have focused on the *Brassicaceae* trichome initiation pathway as an ideal study system because of the knowledge available from previous numerous experiments (Han et al 2022, Pattanaik et al 2014, Wang et al 2019, Yang & Ye 2013), as trichomes have been regarded as an important epidermal structure for plants (Karabourniotis et al 2020, Wang et al 2021b) which are also easily observable when experimenting plant cell differentiation and development (Kryvych et al 2011). This pathway is being examined in the allopolyploid *Pachycladon* genus, which formed from highly diverged diploid progenitors (Joly et al 2009) and has since diversified into ~10 species (Heenan et al 2002, Mirzaei et al 2011). We are focusing on four critical genes in the pathway *TTG1*, *GL1*, *GL3,* and *EGL3* (Pattanaik et al 2014, Tominaga-Wada et al 2011, Wang et al 2019). The current chapter focuses on isolating the two homeologs of each gene from each of four different *Pachycladon* species and assigning them to 'ancestral' pathways. Subsequent chapters focus on the molecular evolution of the genes (Chapter 3) and expression analyses (Chapter 4).

2.5.1 Sequence analyses

The phylogenetic analyses of all four genes showed nearly identical results concerning the placement of *Pachycladon* Type 1 and Type 2 homeologs. For each gene, the Type 1 homeologs formed a monophyletic group (noting that *EGL3* only had a single *Pachycladon* sequence) that was sister to the *Capsella* and *Camelina* orthologs and that grouping was sister to the *Arabidopsis* orthologs, effectively forming a large 'lineage I' group. For *TTG1*, *GL1*, and *GL3*, the *Pachycladon* Type 2 homeologs formed a monophyletic group (noting that there was only a single representative for *GL3)* that was sister to the larger 'lineage I' grouping. Determining whether or not these Type 2 homeologs should be considered as part of the Brassicaceae lineage 1 group will require ortholog sequences from more Brassicaceae species.

The sequence divergence between the two types was found to be in the range of 9.3-11.4%, which is quite high. The estimated divergence time for the progenitors is ~6.4 myr

before the hybridization event (Joly et al 2009). For comparison, the estimated divergence times for *Gossypium* (parental divergence 5.2 myr) (Senchina et al 2003) and *Spartina anglica* (Fortune et al 2007), which have been suggested to represent the greatest nrITS divergence between parental genomes (Chapman & Burke 2007), are younger than *Pachycladon*'s parental divergence time before their hybridization (Behling et al 2020).

The progenitor species of *Pachycladon* are unknown, although multiple studies have suggested that the progenitor more closely related to *A. thaliana* was the paternal progenitor (Hendriks et al 2023, Joly et al 2009). The availability of complete gene sequences from *A. thaliana* meant that isolating the Type 1 sequences from *Pachycladon* was generally more straightforward than isolating the Type 2 homeologs.

Several polymorphisms were observed in individual homeologs that are expected to be highly disruptive to gene/protein function. This will be discussed in depth in the next chapter in terms of in silico translations and predicting changes to gene/protein function. For example, the *P. fastigiatum GL1* Type 2 homeolog has a 5 bp deletion in exon 1 that causes a frameshift mutation and early stop codon. Interestingly, the closely related *P. enysii* has an independent mutation in its *GL1* Type 2 sequence, an 18 bp deletion in the known transcriptional activation domain, near the 3' end that would also be expected to significantly affect function.

GL3 and *EGL3* genes belong to the bHLH family, which is known to be one of the largest families of transcription factors in *A. thaliana* (Carretero-Paulet et al 2010, Hao et al 2021). Four bHLH proteins, GL3, EGL3, TT8, and AtMYC1, are involved in *Arabidopsis thaliana* trichome morphogenesis, performing overlapping, yet divergent roles (Balkunde et al 2010, Symonds et al 2011, Zhang et al 2003, Zhao et al 2012b, Zhao et al 2008). The roles of GL3 and EGL3 are important for trichome branching and endoreduplication (section 1.2.1.2) (Payne et al 2000, Wang et al 2019). Further, *TT8* is involved in leaf margin trichome formation (Maes et al 2008). Also, *AtMYC1* mutants have shown a reduced number of trichomes, a defect that was successfully complemented by *GL3* and *EGL3* (Zhao et al 2012b). The sequence similarity among the bHLHs and the relatively low coverage of any one gene from the MiSeq runs made confident assemblies of the *Pachycladon* sequences for these genes

more difficult. In the future, working directly with cDNAs and taking a PCR/Sanger sequencing approach would likely be more fruitful. In the current study, time ran short. However, of the sequences (some incomplete) that were isolated, several interesting mutations were identified. The potential mutational effects of *GL3* and *EGL3* sequences will be discussed in depth in Chapter 3.

2.5.2 Phylogenetic analysis of the genes to reconstruct progenitor pathways

Inconsistencies between the gene trees produced here and recent Brassicaceae phylogenies (Hendriks et al 2023, Nikolov et al 2019) can easily be ascribed to differences in the number and breadth of taxa represented and the number of loci utilized. As the focus here was on identifying the phylogenetic placement of homeologs of individual genes, our analyses are limited to the gene of interest (in terms of the number of loci used to reconstruct a phylogeny) and to the taxa for which complete coding sequences were available. This approach of using phylogenies to infer the ancestral origin of allopolyploids seems solid and has been often used in the literature (Díaz-Pérez et al 2018, Grover et al 2015, Luo et al 2017, Marcussen et al 2015), but not necessarily to find the origin of gene complexes or pathways, maybe because it is still an understudied area.

An important finding of our data is that the relationships among the homeologs of each gene and the sampled taxa are fairly consistent, as described above. The inference here is that Type 1 homeologs that cluster with lineage 1 all trace back to one of the ancestors of the *Pachycladon* genus and Type 2 sequences that cluster outside of lineage 1 trace back to the other ancestor, proving our hypothesis of the origin of homeologous copies. Such ‘ancestral pathway’ assemblies will then allow for a better context for comparisons in progenitor pathway evolution (Chapter 3) and expression (Chapter 4).

The resulting phylogenetic placements of homeologs for all genes show analogous placements to most of the previous studies (See 1.3.2 for a detailed description of the previous studies done) (Huang et al 2015, Joly et al 2009, Zhao et al 2010), resembling Type 1 homeologs positioned in lineage I and Type 2 homeologs positioned outside lineage I.

Based on Joly et al (2009)'s maternal phylogenetic inference of *Pachycladon* species using the chloroplast gene *rbcL* with 52 other *Brassicaceae* homologs, the gene copy which was located close to the split of the *Arabidopsis* and *Brassica* lineages was inferred as the maternally inherited copy (Joly et al 2009). Hendriks et al (2023)'s most complete Brassicaceae phylogeny to date, with 60 plastome genomes and 265 genera, also places the *Pachycladon* maternal contribution outside of the clade that houses *Arabidopsis*, *Camelina*, *Capsella*, *Turritis* and *Erysimum* groups, similar to the Joly et al (2009) study. These phylogenetic placements best correspond to the placement of Type 2 homeolog copies in the present study. Taken together, we infer the Type 1 homeologs to be of paternal origin and the Type 2 copies to be of maternal origin. Although Mandáková et al (2010)'s study showed *Pachycladon* Type 1 copy in our study is likely the maternal copy using an inferred phylogeny of four chloroplast genes (*rbcL*, *nad4*, *matK*, *ndhF*) with 55 other *Brassicaceae* homologs.

In the sequence analyses and the gene trees (Figure 2.3, Figure 2.4, and Figure 2.5) it was regularly observed that the *P. enysii* and *P. fastigiatum* sequences are very similar. Their coding sequences contained very similar patterns of SNPs and had the highest percent similarity values among all sequences. Although these species are morphologically different from each other and have divergent habitat preferences (*P. enysii* is high alpine and *P. fastigiatum* is mid-alpine), our results match previous reports that the two species are phylogenetically very close (Joly et al 2014). *P. cheesemanii* appears to be the sister group to other species particularly in *TTG1* type 1 *Pachycladon* clade and in previous studies (Heenan & Mitchell 2003, McBreen & Heenan 2006) (although not separated in other type-specific *Pachycladon* clades in the gene trees presented in the results). *P. cheesemanii* has a wide ecological tolerance and is considered to be a generalist and also has wide latitudinal and altitudinal ranges, and it may resemble the ancestral form of the genus in New Zealand, while all other species have distinct and relatively narrow ecological niches (Heenan & Mitchell 2003).

2.6 CONCLUSION

Pachycladon is a young allopolyploid lineage originating in the Pleistocene from significantly diverged *Brassicaceae* parental lineages. Thus, it is an ideal model species to study the molecular evolution of a duplicated genetic pathway. The trichome initiation pathway was chosen as a good candidate system because Brassicaceae trichome morphogenesis is reasonably well-defined. Exploring the origin of the duplicated gene copies in the trichome initiation pathway in *Pachycladon* is the first step in the molecular evolutionary analysis of the pathway. The ancestral origin of four gene homeologs involved in constructing the conserved transactivation complex in the trichome initiation pathway belonging to four *Pachycladon* species was investigated in this chapter. Thus, the initial assessments of Types 1 and 2 are confirmed and each Type seems to represent one of the parental copies. The phylogenetic analysis revealed that the Type 1 copy of the genes belongs to Lineage I in the *Brassicaceae* phylogeny (Nikolov et al 2019), relating closely to *Arabidopsis*, *Camelina,* and *Capsella*, while the Type 2 copy forms a monophyletic clade sister to Lineage I; however, Type 2 may also be part of Lineage 1, just further out. The phylogenetic positions of the two homeologous copies were consistent among the gene trees and in line with the previous experiments of (Huang et al 2015), Joly et al (2009), Zhao et al (2010) and Huang et al (2015). In addition, it is important to note that a lot of change may have taken place since the hybridization event, so these homeologs don't represent the exact copies of the parental genes but do allow their origins to be inferred.

REFERENCES

Al-Shehbaz IA, O'Kane SL, Price RA. 1999. Generic Placement of Species Excluded from Arabidopsis (Brassicaceae). *Novon* 9: 296-307

Alfandary R. 2015. 4 Tips for Efficient Primer Design.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215: 403-10

Balkunde R, Pesch M, Hülskamp M. 2010. Trichome patterning in Arabidopsis thaliana: from genetic to molecular models. *Current topics in developmental biology* 91: 299-321

Behling AH, Shepherd LD, Cox MP. 2020. The importance and prevalence of allopolyploidy in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand* 50: 189-210

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American journal of botany* 93: 607-19

Bennett MD. 2004. Perspectives on polyploidy in plants–ancient and neo. *Biological Journal of the Linnean Society* 82: 411-23

Bernhardt C, Lee MM, Gonzalez A, Zhang F, Lloyd A, Schiefelbein J. 2003. The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root. *Development* 130: 6431-39

Birchler JA. 2012. Genetic consequences of polyploidy in plants In *Polyploidy and genome evolution*, pp. 21-32: Springer

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* 13: 137-44

Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martínez-García JF, Bilbao-Castro JR, Robertson DL. 2010. Genome-Wide Classification and Evolutionary Analysis of the bHLH Family of Transcription Factors in Arabidopsis, Poplar, Rice, Moss, and Algae *Plant Physiology* 153: 1398-412

Chapman MA, Burke JM. 2007. Genetic divergence and hybrid speciation. *Evolution* 61: 1773-80

Collins LJ, Biggs PJ, Voelckel C, Joly S. 2008. An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Inform* 21: 3-14

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20: 591-7

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research* 12: 1075-79

Díaz-Pérez A, López-Álvarez D, Sancho R, Catalán P. 2018. Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus Brachypodium using minimum evolution, coalescence and maximum likelihood approaches. *Molecular Phylogenetics and Evolution* 127: 256-71

Dong Y, Gupta S, Sievers R, Wargent JJ, Wheeler D, et al. 2019. Genome draft of the Arabidopsis relative Pachycladon cheesemanii reveals novel strategies to tolerate New Zealand's high ultraviolet B radiation environment. *BMC Genomics* 20: 838

Doyle JJ, Coate JE. 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180: 1-52

Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. *Focus* 12: 39-40

Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML. 2007. Evolutionary dynamics of Waxy and the origin of hexaploid Spartina species (Poaceae). *Mol Phylogenet Evol* 43: 1040-55

Freeling M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology* 60: 433-53

Freeling M, Scanlon MJ, Fowler JE. 2015. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr Opin Genet Dev* 35: 110-8

Geneious 9.1.8. 2017.
Grant V. 1981. Plant speciation, Columbia University Press
Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, et al. 2015. Re-evaluating the phylogeny of allopolyploid Gossypium L. *Molecular Phylogenetics and Evolution* 92: 45-52
Han G, Li Y, Yang Z, Wang C, Zhang Y, Wang B. 2022. Molecular mechanisms of plant trichome development. *Front Plant Sci* 13: 910228
Hao Y, Zong X, Ren P, Qian Y, Fu A. 2021. Basic Helix-Loop-Helix (bHLH) transcription factors regulate a wide range of functions in Arabidopsis. *International Journal of Molecular Sciences* 22: 7152
Heenan P, Mitchell A, Koch M. 2002. Molecular systematics of the New Zealand Pachycladon (Brassicaceae) complex: generic circumscription and relationships to Arabidopsis sens. lat. and Arabis sens. lat. *New Zealand Journal of Botany* 40: 543-62
Heenan PB, Mitchell AD. 2003. Phylogeny, biogeography and adaptive radiation of *Pachycladon* (Brassicaceae) in the mountains of South Island, New Zealand. *Journal of Biogeography* 30: 1737–49
Hendriks KP, Kiefer C, Al-Shehbaz IA, Bailey CD, van Huysduynen AH, et al. 2023. Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Current Biology*
Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, et al. 2001. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* 29: 102-5
Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, et al. 2015. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. *Molecular Biology and Evolution* 33: 394-412
Joly S, Heenan PB, Lockhart PJ. 2009. A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. *Mol Phylogenet Evol* 51: 365-72
Joly S, Heenan PB, Lockhart PJ. 2013. Species Radiation by Niche Shifts in New Zealand's Rockcresses (*Pachycladon*, Brassicaceae). *Syst. Biol.* 63: 192–202
Joly S, Heenan PB, Lockhart PJ. 2014. Species Radiation by Niche Shifts in New Zealand's Rockcresses (Pachycladon, Brassicaceae). *Systematic Biology* 63: 192-202
Joshi P, Ansari H, Dickson R, Ellison NW, Skema C, Tate JA. 2022. Polyploidy on islands – concerted evolution and gene loss amid chromosomal stasis. *Annals of Botany* 131: 33-44
Karabourniotis G, Liakopoulos G, Nikolopoulos D, Bresta P. 2020. Protective and defensive roles of non-glandular trichomes against multiple stresses: structure–function coordination. *Journal of Forestry Research* 31: 1-12
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* 279: 5048-57
Kryvych S, Kleessen S, Ebert B, Kersten B, Fisahn J. 2011. Proteomics–the key to understanding systems biology of Arabidopsis trichomes. *Phytochemistry* 72: 1061-70
Li Z, McKibben MT, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72: 387-410
Liu B, Xu C, Zhao N, Qi B, Kimatu JN, et al. 2009. Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. *Journal of Genetics and Genomics* 36: 519-28
Luo X, Hu Q, Zhou P, Zhang D, Wang Q, et al. 2017. Chasing ghosts: allopolyploid origin of Oxyria sinensis (Polygonaceae) from its only diploid congener and an unknown ancestor. *Molecular Ecology* 26: 3037-49
Maes L, Inzé D, Goossens A. 2008. Functional Specialization of the TRANSPARENT TESTA GLABRA1 Network Allows Differential Hormonal Control of Laminal and Marginal Trichome Initiation in Arabidopsis Rosette Leaves. *Plant Physiology* 148: 1453-64
Manda´kova´ T, Joly S, Krzywinski M, Mummenhoff K, Lysaka MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell* 22: 2277–90

Mandáková T, Heenan PB, Lysak MA. 2010. Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evol Biol* 10: 367-67

Mandakova T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. *Plant Cell* 22: 2277-90

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal* 91: 3-21

Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus Viola (Violaceae). *Systematic biology* 64: 84-101

Mason-Gamer RJ, Burns MM, Naum M. 2010. Phylogenetic relationships and reticulation among Asian Elymus (Poaceae) allotetraploids: Analyses of three nuclear gene trees. *Molecular Phylogenetics and Evolution* 54: 10-22

McBreen K, Heenan PB. 2006. Phylogenetic relationships of Pachycladon (Brassicaceae) species based on three nuclear and two chloroplast DNA markers. *New Zealand Journal of Botany* 44: 377-86

Mirzaei M, Pascovici D, Keighley T, George I, Voelckel C, et al. 2011. Shotgun proteomic profiling of five species of New Zealand Pachycladon. *Proteomics* 11: 166-71

Mitchell AD, Heenan PB. 2000. Systematic Relationships of New Zealand Endemic Brassicaceae Inferred from nrDNA ITS Sequence Data. *Systematic Botany* 25: 98-105

Moghe GD, Shiu SH. 2014. The causes and molecular consequences of polyploidy in flowering plants. *Ann N Y Acad Sci* 1320: 16-34

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, USA.

Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, et al. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222: 1638-51

Pattanaik S, Patra B, Singh SK, Yuan L. 2014. An overview of the gene regulatory network controlling trichome development in the model plant, Arabidopsis. *Front Plant Sci* 5: 259

Payne CT, Zhang F, Lloyd AM. 2000. GL3 Encodes a bHLH Protein That Regulates Trichome Development in Arabidopsis Through Interaction With GL1 and TTG1. *Genetics* 156: 1349-62

Price JP, Wagner WL. 2004. Speciation in Hawaiian Angiosperm Lineages: Cause, Consequence, and Mode. *Evolution* 58: 2185-200

Price RA, Al-Shehbaz IA, O'Kane SL. 2001. Beringia (Brassicaceae), a New Genus of Arabidopsoid Affinities from Russia and North America. *Novon* 11: 332-36

Schenk JJ. 2021. The Next Generation of Adaptive Radiation Studies in Plants. *International Journal of Plant Sciences* 182: 245-62

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, et al. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* 20: 633-43

Shi X, Zhang C, Ko DK, Chen ZJ. 2015. Genome-Wide Dosage-Dependent and -Independent Regulation Contributes to Gene Expression and Evolutionary Novelty in Plant Polyploids. *Molecular Biology and Evolution* 32: 2351-66

Soltis DE, Buggs RJ, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* 59: 1387-403

Soltis DE, Soltis PS, Tate JA. 2004. Advances in the study of polyploidy since Plant speciation. *New Phytologist* 161: 173-91

Stuessy TF, Crawford DJ. 1998. Chromosomal stasis during speciation in angiosperms of oceanic islands In *Evolution and Speciation of Island Plants*, ed. M Ono, TF Stuessy, pp. 307-24. Cambridge: Cambridge University Press

Symonds VV, Hatlestad G, Lloyd AM. 2011. Natural allelic variation defines a role for ATMYC1: trichome cell fate determination. *PLoS Genet* 7: e1002069

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38: 3022-27

Tominaga-Wada R, Ishida T, Wada T. 2011. New insights into the mechanism of development of Arabidopsis root hairs and trichomes. *International review of cell and molecular biology* 286: 67-106

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* 35: W71-W74

Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* 33: 11-26

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411-24

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. *Science* 290: 2114-17

Wang X, Shen C, Meng P, Tan G, Lv L. 2021. Analysis and review of trichomes in plants. *BMC plant biology* 21: 1-11

Wang Z, Yang Z, Li F. 2019. Updates on molecular mechanisms in the development of branched trichome in Arabidopsis and nonbranched in cotton. *Plant Biotechnology Journal* 17: 1706-22

Weiss-Schneeweiss H, Emadzade K, Jang T-S, Schneeweiss GM. 2013. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenetic and genome research* 140: 137-50

Wendel JF. 2000. Genome evolution in polyploids. *Plant molecular evolution*: 225-49

Yang C, Ye Z. 2013. Trichomes as models for studying plant cell differentiation. *Cellular and molecular life sciences* 70: 1937-48

Yogeeswaran K, Voelckel C, Joly S, Heenan PB. 2011. Pachycladon. 227-49

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18: 821-29

Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. 2003. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. *Development* 130: 4859-69

Zhao B, Liu L, Tan D, Wang J. 2010. Analysis of phylogenetic relationships of Brassicaceae species based on Chs sequences. *Biochemical Systematics and Ecology* 38: 731-39

Zhao H, Wang X, Zhu D, Cui S, Li X, et al. 2012. A Single Amino Acid Substitution in IIIf Subfamily of Basic Helix-Loop-Helix Transcription Factor AtMYC1 Leads to Trichome and Root Hair Patterning Defects by Abolishing Its Interaction with Partner Proteins in Arabidopsis. *Journal of Biological Chemistry* 287: 14109-21

Zhao M, Morohashi K, Hatlestad G, Grotewold E, Lloyd A. 2008. The TTG1-bHLH-MYB complex controls trichome cell fate and patterning through direct targeting of regulatory loci. *Development* 135: 1991-99

A2. APPENDIX

A2.1 Statement of contribution

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Thilini Madara Warusawithana
Name and title of main supervisor:	Dr. V. Vaughan Symonds
In which chapter is the manuscript/published work?	Chapter 2
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

Thilini worked with Dr. Vaughan Symonds to develop the research question and objectives, and to design the experimental setup and data collection methods, including multiple sequencing methods to isolate homeologs. Under the guidance of Dr. Symonds, Thilini also conducted analyses and interpreted the results. Thilini drafted the chapter and incorporated the feedback from Vaughan and co-supervisor Dr. Matthieu Vignes.

Please select one of the following three options:

- ○ The manuscript/published work is published or in press
  Please provide the full reference of the research output:
- ○ The manuscript is currently under review for publication
  Please provide the name of the journal:
- ◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Student's signature:	Thilini Warusawithana Digitally signed by Thilini Warusawithana Date: 2024.10.25 14:50:58 +13'00'	Main supervisor's signature:	Vaughan Symonds Digitally signed by Vaughan Symonds Date: 2024.10.29 11:42:10 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

1

Doctoral Research Committee
July 2022

3. CHAPTER THREE

Analyzing and characterizing the molecular evolution of the trichome initiation pathway across species

3.1 ABSTRACT

Polyploidy or whole genome duplication (WGD) either duplicates the genome of the progenitor or combines the full genome complement of two distinct lineages. Up to this point, molecular work on polyploidy has typically focused on the fates of individual genes or on genome-wide expression patterns. However, an important context that likely affects the fates of individual genes following duplication that has not been widely explored is that of the genetic pathway. The genus *Pachycladon,* being an allotetraploid lineage originating from highly diverged parental lineages, possesses two diverged copies of most genes. Here, we focus on the duplicated well-annotated trichome initiation pathway of *Pachycladon*, which forms a TTG1-GL1-GL3/EGL3 conserved transactivation complex, originally described from the close relative, *Arabidopsis thaliana*. Insights into the molecular evolution of *Pachycladon*'s duplicated trichome initiation pathway were gained by studying sequence divergence, Ka/Ks analyses, and in silico protein analysis of the four genes/proteins of interest. The Type 2 copies of all genes were found to have greater variation than the Type 1 copies. Although tests for natural selection indicated that both copies are largely evolving under the influence of purifying selection, the Type 2 copies are accumulating more mutations. Of particular note was the finding that *Pachycladon fastigiatum* carries a *GL1* Type 2 copy that has a five bp deletion, which causes a frameshift upon translation; this partially explains the lack of trichomes in this species. Other mutations with potential effects on protein function were also identified and are discussed here.

3.2 INTRODUCTION

3.2.1 Polyploidy, doubling of genetic pathways, and subsequent evolution

Whole genome duplications (WGD) copy entire genetic pathways/networks, unlike small-scale segmental duplications, and therefore have been a major driver of lineage diversification, particularly for flowering plants. Following WGD, duplicated networks are rewired, increasing biological complexity, and potentially giving rise to novel traits (De Smet et al 2017, Soltis & Soltis 2016, Van de Peer et al 2009b). WGD events are followed by widescale gene loss through genomic rearrangements (Liang & Schnable 2018, Magadum et al 2013, Wang et al 2016). The fate of duplicated genes, including their retention and expression over time, is a topic of broad importance. Generally, when both copies of a duplicated gene (homeologs) are not quickly lost through genomic rearrangements, they may be retained to allow for dosage compensation or undergo subfunctionalization or neofunctionalization (Blanc & Wolfe 2004a, Conant & Wolfe 2008, De Smet et al 2017). Alternatively, one copy may be maintained and the other lost through small-scale deletions or a process of pseudogenization (McGrath & Lynch 2012, Moghe et al 2014, Yu et al 2020). Many current approaches to studying the fates of duplicated genes remain largely descriptive (e.g., whole genome scans) and lack a critical context: the genetic pathway (Roth et al 2007). As a result, relatively little is known about how the constraints imposed by co-evolved genetic systems, (e.g., pleiotropy, epistasis, and protein: protein interactions) for adaptively important traits impact the fate of gene duplicates.

3.2.2 The duplicated trichome initiation pathway in *Pachycladon*

Pachycladon is an allopolyploid lineage consisting of ~10 species native to New Zealand and Australia that was derived from two highly diverged progenitors. Of the multitude of large-scale Brassicaceae phylogenies published over the last decade (Hendriks et al 2023, Huang et al 2015, Nikolov et al 2019), we are unaware of any that included both *Pachycladon* homeologs of the markers used. As such, a robust determination of the phylogenetic placement(s) of *Pachycladon* does not yet exist. So while the phylogenetic uncertainty persists

and the actual parent species remain unknown (and are likely extinct), based on multiple studies (Joly et al 2009, Mandakova et al 2010) and the results in Chapter 2 of this thesis, one of the progenitors was more closely related to the *Arabidopsis* lineage than the other. With highly diverged homeologs of all genes and a relatively close relationship with the model plant species, *A. thaliana*, *Pachycladon* has been identified as an ideal model plant group in which to study the molecular evolution of duplicated genetic pathway (Joly et al 2009, Yogeeswaran et al 2011).

The pathway of choice here is the trichome initiation pathway that has been described from *A. thaliana* to gain insights into cell fate determination, pattern formation and morphogenesis (Szymanski et al 2000). *Arabidopsis thaliana* trichomes on mature rosette leaves usually have 3-4 branches (Telfer et al 1997) and vary in density within and among natural accessions (Bloomer et al 2012). *Pachycladon* demonstrates variation for trichomes, in the means of morphology, density, and positioning among its species (Yogeeswaran et al 2011). Considering the four *Pachycladon* species under study here, *P. enysii* and *P. cheesemanii* tend to have a higher density of branched trichomes, P. *fastigiatum* typically has glabrous (absence of trichome) leaves rarely with simple trichomes on the leaf margins, and *P. novae-zealandiae* has less densely branched trichomes (Mershon et al 2015, Yogeeswaran et al 2011). Therefore, this pathway in *Pachycladon* provides an excellent study system for the molecular evolution of a duplicated genetic pathway with a clear phenotype context.

3.2.3 The TTG1-GL1-GL3/EGL3 conserved transactivation complex

The evolutionarily conserved *Arabidopsis* trichome initiation transactivation complex comprises proteins/transcription factors belonging to three families. Namely, TTG1, GL1, GL3, and EGL3 transcription factors (TFs) are the critical factors of the Brassicaceae trichome initiation pathway (Payne et al 2000, Zhang et al 2003). While many other proteins play roles in this pathway, here we are focusing on the core of the pathway and those genes/proteins that are best described. Generally, TF binds with DNA directly via DNA binding domains after

recognizing particular DNA sequences or indirectly via protein-protein interactions with a cofactor to initiate the transcription (Jones & Vandepoele 2020, Stracke et al 2001).

The *Arabidopsis* WD40 protein TTG1 contains four WD40 repeat motifs (Figure 3.1a), but not any nuclear localization signal, DNA binding domain (DBD), or transcriptional activation domain (TAD) (Walker et al 1999). Each WD repeat motif is made of characteristic GH (Glycine-Histidine) and WD (Tryptophan-Aspartic acid) amino acids, which are separated by a conserved pattern of hydrophobic and hydrophilic amino acid residues (Walker et al 1999). In the trichome initiation activation complex, the WD40 repeat motifs are important in forming protein-protein interactions with bHLH proteins for nuclear import or retention and acting as a transcriptional co-regulator in downstream processes. Hence, TTG1 acts as both a protein interaction factor in the cytoplasm and a transcriptional co-regulator in the nucleus (Zhao et al 2008). This protein can be regarded as a master regulatory protein that underlies multiple developmental and biochemical *Arabidopsis* pathways as stated in 1.2 and shown in Figure 1.3, as it is involved in the formation of leaf, stem trichomes, root hairs, and the production of seed coat mucilage and anthocyanin pigments (Walker et al 1999). Most recently, a role for TTG1 in flowering time regulation has been described (Paffendorf et al 2020) but that role will not be considered here.

GL1 and MYB23 TFs (MYB23 is also bound in the transactivation complex as stated in 1.2.1.2) belong to the R2R3 MYB family, which is a part of one of the largest transcription factor families, the MYB gene family (Riechmann et al 2000, Romero et al 1998). The MYB DNA-binding domain has conserved helix-turn-helix repeats, and R2R3 MYBs contain two repeats consisting of about 53 amino acids each (Stracke et al 2001). The R2R3 MYB, GL1, possesses two important domains as illustrated in the Figure 3.1b, including a highly conserved N-terminal DNA Binding Domain (DBD) and the variable C-terminal Transcriptional Activation Domain (TAD) modulating TF activation and DNA binding domain accessibility and thereby enhancing the specificity of DNA binding of the transcription factor (Dubos et al 2010, Pireyre & Burow 2015). For binding with bHLHs, a conserved 20 amino acid signature

in the R3 MYB domain is regarded as important (Zimmermann et al 2004). GL1 binds with the bHLHs GL3 and EGL3, but not directly with TTG1 in the transactivation complex.

The bHLHs (GL3 and EGL3) possess a set of 60 conserved amino acid residues in the C-terminal DBD (Figure 3.1c), and more variable N-terminal TAD (Cheng et al 2011, Dombrecht et al 2007). GL3 can form heterodimers with EGL3 (Patra et al 2013, Payne et al 2000, Zhang et al 2003), TTG1, 14 R2R3 MYB proteins (including GL1, WER, and AtMYB23), and 6 R3 MYB proteins and homodimers with itself in *Arabidopsis thaliana* (Zhao et al 2013, Zimmermann et al 2004). Protein-protein interactions of those key components of the transactivation complex have been explored using yeast two-hybrid assays. The N-terminus including the MYB repeats of GL1 and the first 96 amino acids of GL3 are found to be necessary for GL1-GL3 interaction (Payne et al 2000). Also, the GL3 N-terminus binds with TTG1, whereas TTG1 does not interact with GL1 (Payne et al 2000). The C-terminal region containing the bHLH domain is relevant for the formation of homodimers and binds with EGL3 (Payne et al 2000, Zhang et al 2003).

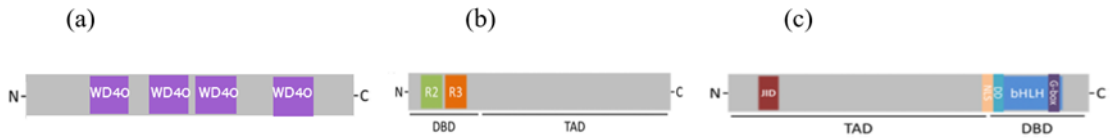


Figure 3.1: Schematic representation of (a) WD40 protein, (b) R2R3 MYB, and (c) bHLH TFs. The WD40 protein contains WD40 repeats without a DNA binding domain (DBD) or transcriptional activation domain (TAD). The R2R3 MYBs have a conserved DBD at the N terminus and a more variable, C-terminal TAD. The bHLHs have a DNA binding motif (G-Box), a dimerization domain (DD), and a nuclear localization signal (NLS) in the bHLH domain at the C terminus and a JAZ interaction domain (JID) in the TAD at the N terminus (Pireyre & Burow 2015).

TTG1, GL1, and GL3 directly target the downstream genes *GL2*, *TTG2*, *CPC*, and *ETC1* in the pathway (Wang & Chen 2008, Zhao et al 2008). The three genes/proteins act as positive regulators in the pathway while R3 single-repeat MYBs negatively regulate the pathway to establish trichome spacing on the leaf surface. TTG1 and R3 negative regulators are mobile between trichome initial cells and surrounding cells that do not produce trichomes

(Pesch & Hülskamp 2009). Hence, paradoxically, TTG1 seems to exhibit both positive and negative regulation in trichome development (Han et al 2022, Larkin et al 1999, Pesch & Hülskamp 2009, Schnittger et al 1999, Wang et al 2022a). Recently, *Arabidopsis* trichome initiation has been explored more in the aspects of chromatin, post-transcriptional, and post-translational levels (Han et al 2022, Wang et al 2022a). Still, they are not discussed here, as our primary focus is on the transcriptional regulatory complex in the trichome initiation.

3.2.4 Paralogs in the transactivation complex and their functional redundancy

The genes in the trichome initiation complex have several paralogs that are functionally diversified. While *GL1* is the master gene specialized for trichome initiation, *MYB23* regulates trichome branching and trichome initiation at leaf edges (Kirik et al 2005, Lee & Schiefelbein 1999, Oppenheimer et al 1991), *MYB5* (along with *MYB23*) controls trichome branching on leaves and trichome morphogenesis on stems in *Arabidopsis* (Gonzalez et al 2009, Li et al 2009b) and *MYB82* also regulates trichome initiation (Liang et al 2014, Zhang et al 2019b), although its specific function is still unknown. *GL1* and *MYB23* share 63% amino acid residue similarity and 92% MYB domain similarity (Kirik et al 2001). Another paralog of *GL1* in *A. thaliana* is *WER*, which is involved in root hair development. *WER* has 57% sequence identity with *GL1*, including changes in cis-regulatory sequences responsible for functional diversification (Lee & Schiefelbein 2001), although the genes can complement one another and mutually recover mutant phenotypes (Lee & Schiefelbein 2001). The roles of GL3 and EGL3 are partially redundant in sharing some similar targets (section 1.2.1.2) (Morohashi et al 2007). Both GL3 and EGL3 do not require dimerization to each other to bind to *GL2* and *TTG2* (Zhang et al 2003), the downstream genes activated by the transactivation complex in the pathway.

3.2.5 Functional divergence of duplicates in gene regulatory network evolution

Genetic pathways represent the interplay between TFs and target genes, and they can be altered as a result of molecular evolution. Sequence changes within the cis-regulatory region of a gene where TFs bind (*cis*-effects) and alterations to the copy number of TFs and their

binding preferences (*trans*-effects) can change TF binding behavior in gene regulatory networks (Jones & Vandepoele 2020). A summary of these modes of network evolution is well elucidated in Figure 3.2, extracted from the literature (Jones & Vandepoele 2020). Unlike *cis*-effects, transcription factors often bind to many targets and therefore trans-effects may be under stronger selection pressure (Signor & Nuzhdin 2018). However, the alternative assumption is that duplicates tend to interact with different partners, with no shared protein-protein interactions, evidenced by duplicates having no significant difference in the number of protein-protein interactions (Carretero-Paulet & Fares 2012, Guo et al 2013). As outlined in Figure 3.2, there are four potential evolutionary outcomes of TF duplication in the aspect of interactions with the target DNA, such as subfunctionalization, redundancy, neofunctionalization, and fractionation. Duplicates tend to be retained in other outcomes, except in biased fractionation where one of the duplicates is lost. For example, De Smet et al (2017)'s experiment using 92 homoeologous gene pairs in *A. thaliana* revealed tissue-specific gene expression patterns following WGD, adapting subfunctions in growth/development and stress response, and Schiessl showed subfunctionalization of flowering time genes in *Brassica napus* (Schiessl 2020). Comparatively, examples of neofunctionalization are rare because of the difficulties interpreting the ancestral functions/pathways and the roles of ancestral genes. Recently Bowman et al (2019) presented multiple examples of neofunctionalized components that arose in the evolution of land plant hormone signaling pathways. Neofunctionalization is expected to be rare compared to subfunctionalization because positive selection dominates (1.1.3.2). Li *et al* have shown fractionation of one gene copy of the duplicated genes in the glycolysis IV pathway in maize (Li et al 2016). These are a few of many studies examining the functional divergence of gene duplicates.

Figure 3.2: Modes of evolutionary changes in TF-DNA binding after genome duplication – the figure is extracted from Jones and Vandepoele (2020). (a) *Cis*-regulatory changes, such as sequence changes within a binding site, that happen in a single target gene, result in loss or gain of TF binding, or modulation of TF binding affinity. *Trans*-regulatory changes are the result of changes to the TF, which can cause changes such as loss of DNA binding and change in DNA binding preference (b) After TF duplication there are four potential evolutionary outcomes: i) The regulatory interactions of the ancestral TF are partitioned between the two duplicates (subfunctionalization), ii) both duplicates retain the connections of the ancestral TF (redundancy), iii) one duplicate retains the connections of the ancestral TF while the other evolves to have a novel set of target genes (neofunctionalization), and iv) one of the duplicates is lost (not shown in the figure). (c) After TF duplication, the potential changes of the target gene interactions are, partitioned regulatory interactions (subfunctionalization), persistence of ancestral interactions (redundancy), and novel regulatory interactions (neofunctionalization) (Jones & Vandepoele 2020).

3.2.6 Mutations in protein-coding regions

The primary cause of structural and functional divergence of genetic pathways is mutations. One type of protein-coding sequence evolution is synonymous change (a nucleotide substitution that does not change the corresponding amino acid in the protein) and sometimes it can be nonsynonymous (a nucleotide substitution that changes the corresponding amino acid in the protein) and they are very important in describing the evolutionary dynamics of coding sequences (Fay & Wu 2001, Li et al 2009a, Zhang et al 2006). Nonsynonymous mutations

could change the protein function and hence are important in evolution whereas synonymous substitutions tend to be evolutionarily neutral, although exceptions exist (Bae & Coller 2022, Presnyak et al 2015).

Comparing the frequencies of substitutions as Ka, the number of nonsynonymous substitutions per nonsynonymous site, and Ks, the number of synonymous substitutions per synonymous site, gives a measure of evolutionary change in a given gene sequence (Nekrutenko et al 2002) and allows for inferences regarding forms of natural selection acting on a gene or region of a gene. When Ka equals Ks (Ka/Ks = 1), that indicates a neutral evolution, whereas Ka being less than Ks (Ka/Ks < 1) indicates negative (purifying) selection and Ka being greater than Ks (Ka/Ks > 1) shows positive (diversifying) selection (Zhang et al 2006). Generally, for most protein-coding sequences, nonsynonymous substitutions occur less frequently than synonymous substitutions (Ka < Ks) (Hurst 2002, Makałowski & Boguski 1998). A recent study of Qiao et al (2019) researched the nature of evolutionary forces acting on duplicated genes derived from different modes of duplication methods and showed that the genes derived from WGD are more conserved reported with smaller Ka/Ks ratios and most of the WGD gene pairs in *Arabidopsis* experienced purifying selection (Ka/Ks < 1). The results suggested that WGD-derived gene pairs experienced long-term purifying selection compared to proximal and tandem duplicates (Chapman et al 2006, Roulin et al 2013).

The impact of nonsynonymous amino acid substitutions depends on the particular type of amino acid substituted. Mutagenesis has been widely used to identify the amino acids important for specific motifs required for specialized function or ligand binding. In a recent study analyzing 34,373 mutations in 14 proteins (Gray et al 2017), methionine was stated as the most tolerated substitution and proline as the least tolerated substitution since proline disturbs the conformation of the polypeptide backbone and hydrogen binding. Replacing tryptophan was identified to be highly deleterious also (Dayhoff & Eck 1972, Grantham 1974, Henikoff & Henikoff 1992). The impact of the mutation of 20 amino acids is shown in Figure 3.3, extracted from Gray and coworkers' study, which reflects the median mutational effect score scaled from

0-1, relative to the wild type, where a score of one means wild-type-like activity and a score of zero means destructive to wild type activity. One of the two major classes of amino acids is large hydrophobic amino acids and the other one is polar and charged amino acids. If the wild type and the substituted amino acid pair belong in the same physicochemical type, there will be less effect on function (Gray et al 2017). The study also revealed the secondary structure context of a protein, whether it is an α helix, β sheet or turn does have a significant influence on the mutational tolerance. In addition, cysteine residues are important in forming disulfide bonds in proteins which are important structural elements building protein conformations (Srinivasan et al 1990).

Figure 3.3: The median mutational effect score of amino acid substitutions, extracted from (Gray et al 2017).

3.2.7 *Pachycladon* divergence

When investigating the molecular evolution of the trichome initiation pathway consisting of the important proteins and transcription factors explained above, we suggest three key stages of diversification during *Pachycladon* evolution: 1) the divergence of the two progenitor lineages of *Pachycladon* from a common ancestor, 2) after hybridization between the progenitors and the whole genome duplication event (it is unknown for what period, if any, the hybridized lineage existed before the WGD event occurred, so these are lumped together here), and 3) along individual lineages during the diversification and speciation within *Pachycladon*. Mutations that occurred in the first and second stages would be expected to be largely shared among all extant *Pachycladon* species, whereas the mutations that occurred in individual *Pachycladon* lineages following speciation are not expected to be common among all species. The latter might be specific to individual species or a subset of species in a monophyletic clade. Among the drivers in the above stages, whole genome duplication is likely

to be the most effective force leading to major evolutionary changes in the pathway (Jiao et al 2011, Magadum et al 2013, Sémon & Wolfe 2007). The resulting divergences in the coding sequence, regulatory motif, and expression are expected to increase with the age of the duplicates (Li et al 2005, Qiao et al 2019).

3.2.8 Molecular evolution of the *Pachycladon* trichome initiation pathway

Although numerous studies have been done on the molecular evolution of single genes and in a whole genome context (Clark & Donoghue 2018, Del Pozo & Ramirez-Parra 2015, Sankoff & Zheng 2018), few studies have explored the molecular evolution of genetic pathways following duplication (Kaltenegger et al 2018, Nougué et al 2014, Roth et al 2007). The gene regulatory network level of analysis provides for an examination of the potential impacts of functional interdependency on the molecular evolution of duplicated genes. The highly diverged copies of the trichome initiation pathway in *Pachycladon* make for an ideal system to this end.

With more large deletions and deleterious mutations leading to premature stop codons identified in Type 2 genes/proteins in the sequence analysis in Chapter 2, we hypothesize the Type 2 copies have more sequence divergence, and purifying selection acting on them. Hence, the Type 1 gene copies may be experiencing subgenome bias while Type 2 copies may be experiencing fractionation to maintain the gene dosage balance of the pathway. Nucleotide sequence analysis of duplicated gene copies of the master protein TTG1 suggested that both copies may be intact without any deleterious mutations compared to other genes. Hence, we also hypothesize both copies of TTG1 are preferentially retained because of its highly pleiotropic nature.

3.3 MATERIALS AND METHODS

The following analyses were performed on *TTG1* and *GL1* homeologous sequence alignments from *P. enysii*, *P. fastigiatum*, *P. novae-zealandiae*, and *P. cheesemanii*. *GL3* and *EGL3* homeologous sequences from the same species were not included in the analyses because the data sets for these genes were not complete with gaps in the sequencing, as outlined in Chapter 2. Obtaining complete sequences from these genes and validating them were complex as they are long genes with many paralogs that have high sequence similarities.

3.3.1 Haplotype network construction

For each of the two genes of interest for which complete sequence sets were available (*GL1* and *TTG1)*, the software PopART (Leigh & Bryant 2015) was used to produce a haplotype network for both homeologs from all species. The input for each gene was a sequence alignment made in Geneious and exported in Nexus format. Haplotype networks were generated using the Median Joining network inference method.

3.3.2 DNA polymorphism analysis

Pachycladon multiple sequence alignments that included both Types from all four species for each gene of interest were generated in Geneious (Geneious 9.1.8 2017) (2.3.7.3). A FASTA formatted input file was produced for each alignment to conduct DNA polymorphism analyses using DNASP v.6 (Rozas et al 2017). To analyze the sequence divergence between Type 1 and Type 2 homeologs, two sequence subsets were defined as 'Type_1' and 'Type_2' (using 'Define sequence sets' in the data menu), corresponding to the four copies (one from each species) of each homeolog. Then the nucleotide diversity (Pi) was calculated using a window size of 100 bp and a step size of 25 bp. The positions of functional domains were determined using Uniprot (Coudert et al 2023) and the literature and nucleotide diversity was plotted to show homeolog divergence across the length of each gene.

3.3.3 DNA divergence analysis in synonymous and nonsynonymous sites

In DNASP, the Type 1 sequence set of *TTG1* and *GL1* genes was defined as intraspecific data, and the Type 2 sequence set was defined as interspecific data. Nucleotide

divergence between these two groups was measured for synonymous and nonsynonymous sites using the analysis option 'polymorphism and divergence' with a sliding window of 50 bp and step size of 9 bp, and the results were graphed in Canva (https://www.canva.com/graphs/). The departure of sequence groups from the neutral model of evolution was tested by calculating Tajima's D (Tajima 1989) and Fu and Li's D* and F* (Fu & Li 1993) using the total number of mutations.

3.3.4 Protein analysis

Amino acid alignments of *Pachycladon* species for *TTG1* and *GL1* genes in Geneious (Geneious 9.1.8 2017) were uploaded in FASTA format into the batch processing of the Phyre2 web portal for protein modeling, prediction, and analysis (Kelley et al 2015). The tool uses protein homology detection to build 3D models, predict ligand binding sites, and analyze the effect of amino acid variants. First, the query sequences were scanned against sequences with protein sequence databases with HHblits to build a multiple homologous sequence alignment, which is then used to predict secondary structure with PSIPRED. Next, both the alignment and secondary structure prediction are combined to build up a hidden Markov model, which is scanned against a database of hidden Markov models of proteins of known structures to construct crude backbone-only models. Indels in these backbones are corrected by loop modeling and finally, amino acid side chains are added to generate the Phyre2 model.

Given this model, the Phyre Investigator tool was used to perform in-depth analyses on the model quality; i.e. ProQ2 quality assessment (to predict the quality of protein models) (Ray et al 2012), clashes (to investigate if any of the residues laying too close to others cause clashes), rotamers (to identify sidechains that may not have been modeled ideally, which might affect the backbone or underlying alignment in the region), Ramachandran analysis (to analyze whether the residues in the model lie in favorable, allowed or disallowed regions of the Ramachandran plot) (Ramachandran et al 1963), alignment confidence (to measure the reliability of the pairwise query-template alignment as reported by HHsearch) and disorder prediction (to analyze the dynamically flexible regions in the model, which are static in a

solution that may hinder successful protein crystallization) (Ward et al 2004). The following in-depth analyses were performed to predict the functional attributes of the model: conservation using Jensen-Shannon Divergence (Capra & Singh 2007), pocket detection by fpocket2 (Schmidtke et al 2010) to predict active sites, interface detection to predict protein-protein interaction sites and prediction of mutational effects by SuSPect (Yates et al 2014). Conservation analysis predicts the residues important for the proper structure and function of the protein, whereas others can be more flexible to be replaced. PI-site interface residues are predicted based on the template and other homologs, using the PI-Site protein-protein interface database. When analyzing the proteins in the above aspects, Type-specific/species-specific SNPs and important domains of the query proteins were considered. The GO annotations of the top template in PDB were compared with GO molecular function annotations of the *A. thaliana* homologs to understand the overall functional effects of *Pachycladon* amino acid mutations based on the close genetic relationship of *Pachycladon* with *A. thaliana*. Thus, any functional divergences of the homeologs are predicted.

For *Pachycladon GL3* and *EGL3* sequences, complete coding sequences from start to stop were not available for every species, i. e. for *GL3* Type 1, all complete homeologs were available from the four species, but for Type 2, only *P. cheesemanii* has the complete sequence. *P. novae-zealandiae* has a nearly complete sequence but missing 36 bp in the second exon. For *P. enysii* and P. *fastigiatum*, only one-third of sequences are available from the start codon which ends at the start of exon 5. For *EGL3*, complete coding sequences from start to stop are only found in *P. cheesemanii* both homeologs. Other homeologous copies from the rest of the species were almost complete except few bp regions missing in different places throughout the sequences. Based on these sequences available for the two bHLH genes, they could not be fed for a complete protein analysis due to lack of complete sequences, but they were analyzed for obvious indels for potential frameshift mutations and potential destructive amino acid substitutions such as proline, tryptophan, and cysteine.

Additionally, protein sequences were carefully analyzed in Geneious to identify type-specific and species-specific polymorphisms to predict the evolutionary stage of the origins of polymorphisms and to detect any patterns of homeolog function change related to the parental origins. Empirical results from *A. thaliana* from the literature were used to deduce predictions on the retention or loss of *Pachycladon* homeologs related to functional attributes such as pleiotropy, epistasis, and functional overlap.

3.4 RESULTS

3.4.1 Haplotype network construction

Haplotype networks were constructed for *TTG1* and *GL1* to deduce the relationships between the eight copies (haplotypes) of each gene. The *TTG1* sequences were split into two major haplogroups based on a large number of polymorphisms (87 mutations) that are fixed differences between the homeologs (Figure 3.4a). Seven haplotypes were identified for *TTG1* as *P. enysii* and *P. fastigiatum* have identical sequences for the Type 1 homeolog. For both homeologs, the close evolutionary relationship between *P. enysii* and *P. fastigiatum* is reflected. For *GL1*, eight different haplotypes were identified, again split into the two homeologous groups by a large number of polymorphisms (44 mutations) (Figure 3.4b). Here too, the close evolutionary relationship between *P. enysii* and *P. fastigiatum* is observed for both homeologs.

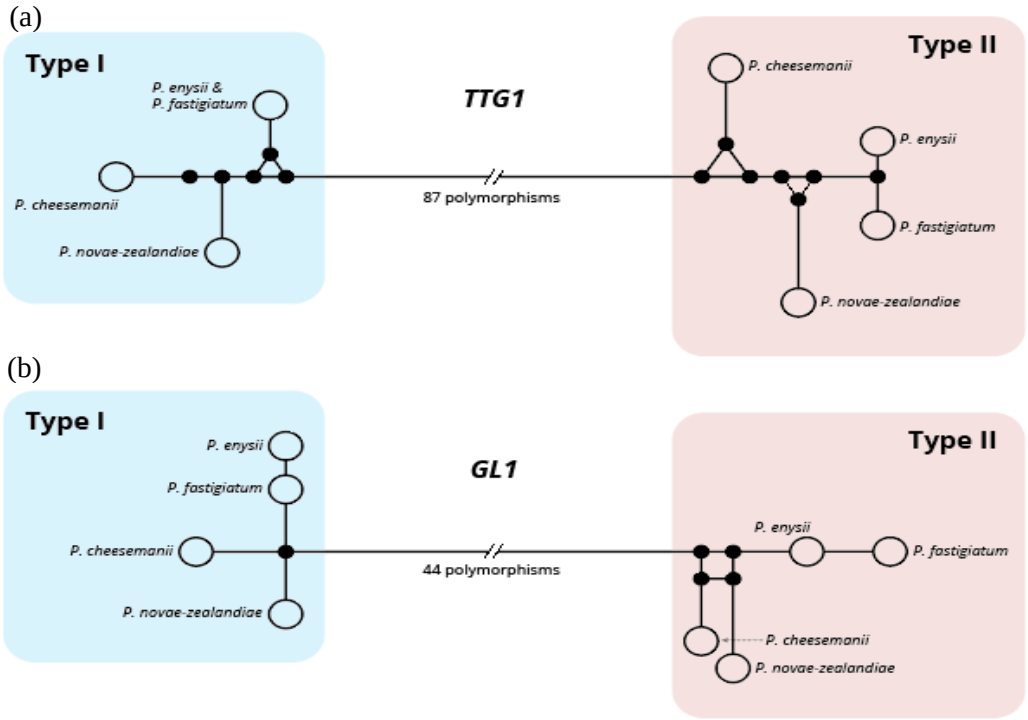


Figure 3.4: Median-joining haplotype network of (a) *TTG1* and (b) *GL1* coding DNA sequences. Sampled haplotypes are represented by open circles and hypothetical unsampled haplotypes necessary to complete the network are represented by smaller closed circles. Branch lengths reflect the number of polymorphisms separating haplotypes.

3.4.2 DNA polymorphism analysis

The extent of DNA divergence between Type 1 and Type 2 homoeologous sequences was analyzed here. The below graphs in Figure 3.5 show the nucleotide diversity (Pi) of Type 1 homeologs and Type 2 homeologs.

The sliding window analysis of nucleotide diversity (Pi) was performed across the *TTG1* coding sequence (Figure 3.5a). The results identified relatively low nucleotide diversity in the Type 1 population upstream of WD1 and in a region of 200 bp from the second half of WD2 across WD3. The Type 1 sequences have relatively high diversity peaks within WD1, at the beginning of WD2, and just upstream of WD4. In contrast, the Type 2 sequences have regions of low variation in WD1, WD2, C-terminal end of WD3, and in between WD3 and WD4 motifs, and higher peaks of variation in the other areas. Interestingly, the two haplotypes show contrasting patterns of nucleotide diversity; when one shows high diversity, the other is more conserved in the same region. However, it seems the variation gradually decreases from the C-terminal half of WD3 and across WD4 in both cases. Overall, the *TTG1* Type 1 sequences exhibit lower nucleotide variation than the Type 2 sequences, i.e., the overall nucleotide diversity value of the Type 1 *TTG1* is 0.00648, and that of the Type 2 *TTG1* is 0.01053. Still, both are quite low.

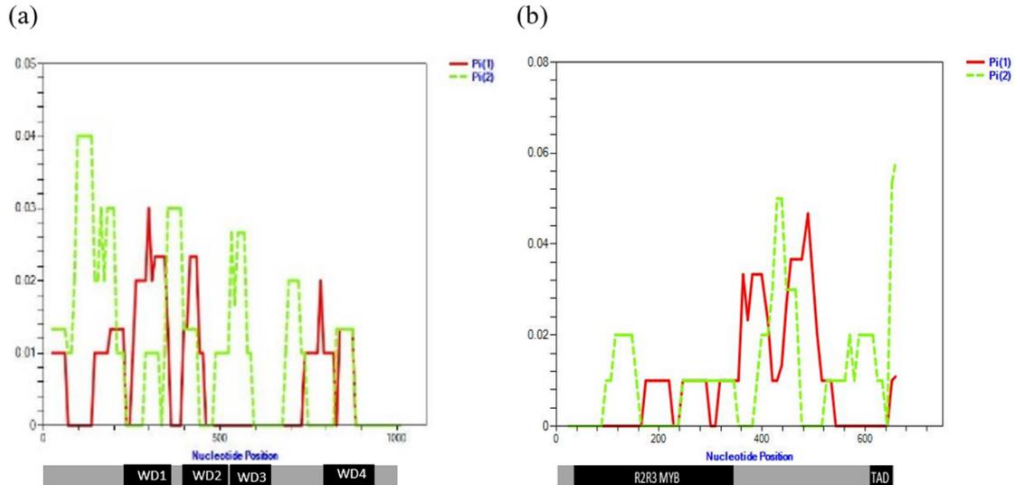


Figure 3.5: The nucleotide diversity (Pi) of *Pachycladon* Type 1 homeologs (red line) and Type 2 homeologs (green dashed line) against the nucleotide position of (a) *TTG1* and (b) *GL1* nucleotide sequences. Sites with alignment gaps were also counted in the sliding window length of 100 bp and step size of 25 bp. The schematic of genes below each graph shows the positions of known functional motifs aligned with the nucleotide position of the graph.

When the sliding window analysis of Pi was performed across the *GL1* coding sequence (Figure 3.5b), low variation was identified in both populations throughout the first half of the gene, where the R2R3 MYB domain is located and again near the 3' end where TAD is located, showing that those functional domains are rather conserved. The region between the two known functional domains and the 3' end shows high peaks of nucleotide variation for both Types. This reflects the high frequency of polymorphism identified in the third exon of both Types. Overall, both Type 1 and Type 2 sequences have fairly similar values for nucleotide diversity, with the Type 1 population being a little lower than 0.009 and the Type 2 population at 0.013.

3.4.3 Analysis of molecular evolution using synonymous and nonsynonymous sites

Sliding window analyses for the ratio of nonsynonymous nucleotide diversity to synonymous nucleotide diversity Pi(a)/Pi(s) and the ratio of nonsynonymous nucleotide divergence to synonymous nucleotide divergence Ka/Ks were performed. However, sites (or codons) with alignment gaps or missing data were not used because of software limitations. Two data sets were defined as a Type 1 homeologous sequence set and a Type 2 homeologous sequence set for *TTG1* and *GL1* genes, but *GL3* and *EGL3* sequences were not included because of a lack of complete sequences. Here we were interested in analyzing the divergence between the two homelogous Types. The ratio of Ka/Ks plotted against the nucleotide position of each gene is shown in Figure 3.6.

Based on these plots, the Ka/Ks ratio is much less than 1 across most of both *TTG1* and *GL1*, except for one or two peaks for each gene. Generally, in coding sequences, nonsynonymous substitutions occur less frequently than synonymous substitutions making Ka/Ks < 1 reflecting purifying selection. Values greater than one are often interpreted as revealing regions of diversifying selection. Only one narrow region in *TTG1* and three regions in *GL1* reveal peaks >1, suggesting those regions have experienced diversifying selection. However, none of them is located in the known functional motifs of the genes. Across the entire coding region, the analysis of Type 1 vs Type 2 for *TTG1* resulted in a Ka/Ks ratio of

0.047, and for *GL1* resulted in a Ka/Ks ratio of 0.460 suggesting both genes are under purifying selection but that *GL1* homeologs have diverged more than the *TTG1* homeologs in terms of nonsynonymous change rates vs synonymous change rates. It is worth noting that sites with alignment gaps were not counted in the sliding window analysis owing to the limitations of the software.

An approach to evaluating the potential departure from neutral expectation for the Types of each gene of interest utilized Tajima's D and Fu and Li's D* and F* calculations for Type 1 and Type 2 homeolog groups. The results are presented in Table 3.1. For all comparisons, statistical tests were not significant (P> 0.10) suggesting that all mutations in the genes are under a neutral model of evolution (Kimura 1983). This interpretation contrasts with the Ka/Ks analysis which suggested the genes are under purifying selection.

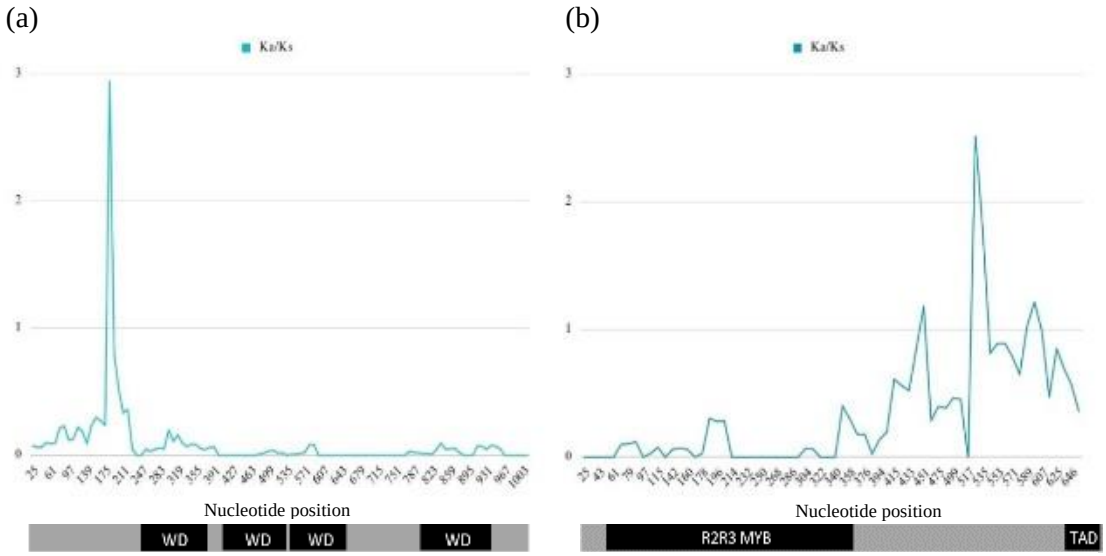


Figure 3.6: ratio of nonsynonymous nucleotide divergence to synonymous nucleotide divergence (Ka/Ks) against the nucleotide positions of *Pachycladon* genes (a) *TTG1* and (b) *GL1* in the sliding window 50 with step size 9. The schematics of genes' functional motifs are aligned with the nucleotide positions below each graph.

Table 3.1: Molecular evolution tests for *Pachycladon* homeolog datasets

***Pachycladon* homeolog group**	**Tajima's D**	**Fu and Li's D***	**Fu and Li's F***
TTG1 Type 1	0.18677 P > 0.10	0.18677 P > 0.10	0.18886 P > 0.10
TTG1 Type 2	-0.54105 P > 0.10	-0.38443 P > 0.10	-0.42784 P > 0.10
GL1 Type 1	0.00000 P > 0.10	0.00000 P > 0.10	0.00000 P > 0.10
GL1 Type 2	-0.07075 P > 0.10	-0.07075 P > 0.10	-0.07262 P > 0.10

3.4.4 Protein analysis

3.4.4.1 TTG1 protein analysis of significant amino acid variations

Phyre2 predicted the top protein model for each *Pachycladon* homeolog based on the confidence, sequence identity, and template information (protein database molecule type and function). However, the predicted top PDB protein molecule types may not be related to the query TTG1 proteins because the TTG1 protein has not yet been modeled in PDB, and so the query sequences were modeled based on the available best matching PDB proteins. The percent identity of the model with the template was low (21-26%), and there was a lot of variation among the predicted models for TTG1 homeologs from different *Pachycladon* species, therefore, the predicted models were not considered in the analysis.

When all the TTG1 sequences were scanned against the protein sequence database with Hhblits, the resulting multiple-sequence alignments up to 1000 homologous sequences had low E-values (<0.001) which is a representation of high confidence good alignments, which in turn generate an accurate secondary structure prediction.

All polymorphisms in the *Pachycladon* TTG1 homeologs were examined carefully relative to the *A. thaliana* reference. The substitutions under the same physicochemical category (but this is not always the case because sometimes substitutions from the same category could have a large effect, e.g., Tryptophan (W), Tyrosine (Y) and Phenylalanine (F) substitutions in the hydrophobic group), substitutions laying outside the functional domains (unless considered significant) and the substitutions in the regions of poor model quality were not focused on here assuming they are not potentially disruptive variants or because of a lack of information. The alignment of *Pachycladon* and *A. thaliana* TTG1 sequences is shown in Figure 3.7 with functional domains annotated. The three amino acid insertion 'AAN' in all Type 1 sequences in the 34th position is not expected to have any disruptive functional consequences because alanine is a well-tolerated hydrophobic uncharged amino acid, and 'N-asparagine' is a polar uncharged amino acid, but not very disturbing (Kelley et al 2015)

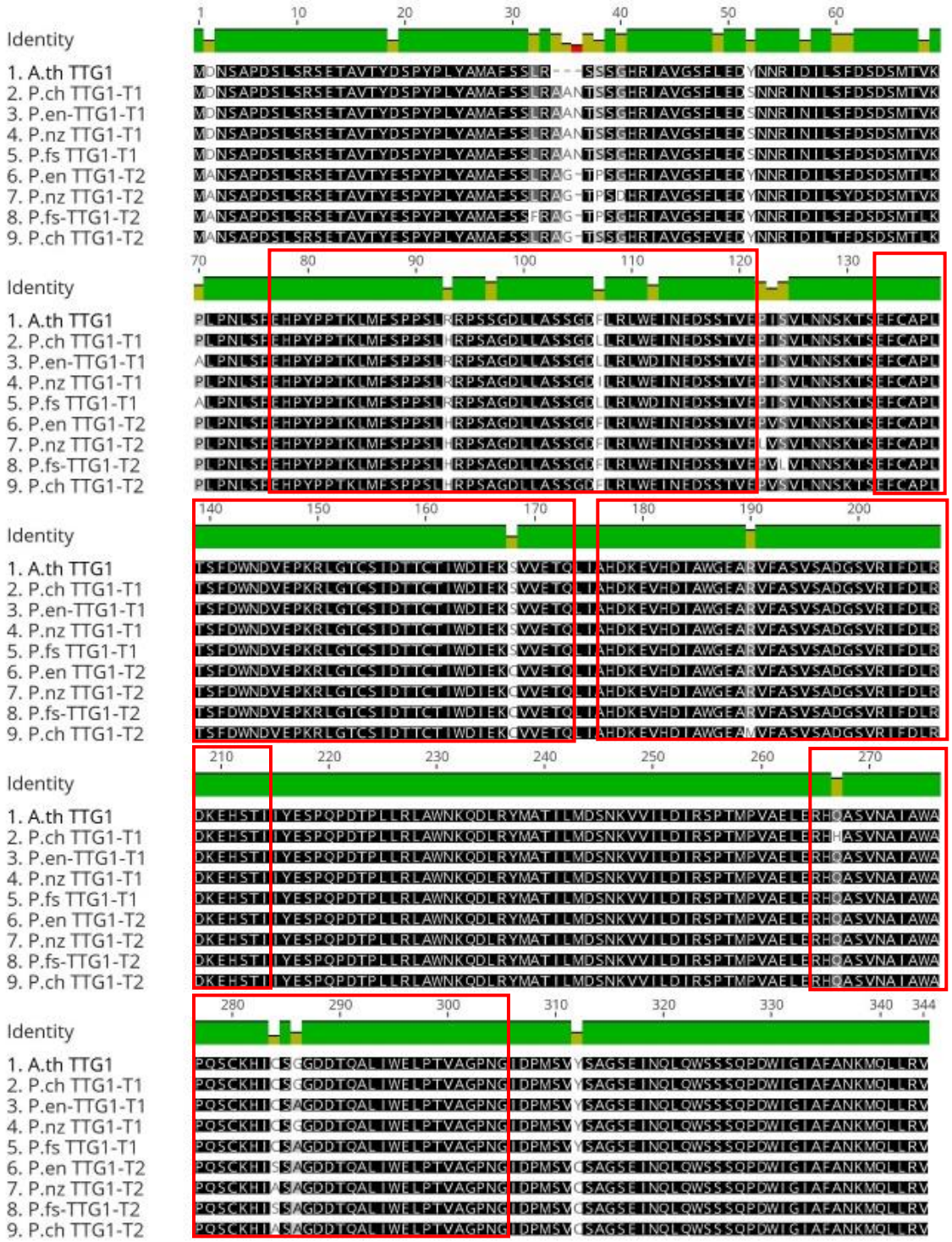


Figure 3.7: Alignment of *A. thaliana* and *Pachycladon* species TTG1 amino acid sequences. Outlined in red are the functional motifs based on the *A. thaliana* sequence; residual positions 77-121 WD1 domain, 133-173 WD2 domain, 176-241 WD3 domain and 265-305 WD4 domain. Positional information on domains was extracted from Uniprot (Coudert et al 2023) and previous studies (Walker et al 1999). The degree of sequence identity is shown on the top of the alignment. The species names are abbreviated here such as A. th - *A. thaliana*, P. ch - *P. cheesemanii*, P. en - *P. enysii*, P. nz - *P. novae-zealandiae* and P. fs - *P. fastigiatum*.

Alanine-glycine insertion is similarly well conserved for the positions and even allowed in the Ramachandran plot for *P. enysii* and *P. fastigiatum*. Low conserved residues in the model are unlikely functionally important positions in Type 2 sequences and are also unlikely to affect the function because they are not regarded as well-conserved, and their mutational sensitivity was found to be low in Phyre2 investigator analysis. The substitutions found to be unfavorable relative to amino acid residues in *A. thaliana* are, P67A, R190M and C284S and only C284S is found to be sitting in a functional domain 'WD4' (Table 3.3). Explaining these mutations, proline (P) in the 67th position is replaced by hydrophobic uncharged alanine (A) in Type 1 *P. enysii* and *P. fastigiatum*, positively charged arginine (R) is replaced by hydrophobic uncharged methionine (M) in *P. cheesemanii* Type 2 at 190th position and the special amino acid cysteine (C) is replaced by polar uncharged serine (S) in Type 2 *P. enysii* and *P. fastigiatum* at 284th position, Cysteine is effective in making disulfide bonds in proteins (Gray et al 2017). But none of these mutations' sensitivity was regarded as high by Phyre2 suggesting they are unlikely to affect the protein function.

Generally, TTG1 residues that impact functionally are well-conserved residues with high mutational sensitivity, and sometimes even present in or close to active sites. Those significant residues are listed in Table 3.2 below after in-depth model quality and functional analysis using Phyre2 investigator. These amino acids are most preferred per position and their mutational sensitivity is high no other substitutions are allowed in the Phyre2 mutational analysis graph. Importantly, all those highly decisive residues have not been replaced by other residues as shown in the graph. Most of them are located in or around active sites as shown in the pocket detection. Protein interaction interface and disorders have not been predicted for the dataset.

Table 3.2: Significant amino acid mutations and functionally important amino acids in the *Pachycladon* proteins TTG1 and GL1

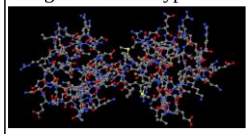
Protein	Species	Position	Domain/SNP/Indel relative to the reference *A. thaliana*	Residue	Functional analysis				Quality analysis of 3D structure						Comments
					Conservation	Mutational sensitivity	Pocket detection	PI-Site interface	Alignment confidence	ProQ2 quality assessment	Ramachandran analysis	Clashes	Rotamers	Disorders	
TTG1 Type 1	*P. en, P. fs*	67	SNP	Proline (P) to Alanine (A)	Low	Low			Moderate	Moderate	Allowed	No			P is more favourable and A is unfavourable per position
TTG1 Type 1 & Type 2	*P. en, P. fs, P.nz , P.ch*	78	WD1	Histidine (H)	Moderate	Moderate			High	Moderate		Yes	Good		
	P. en, P. fs, P.nz , P.ch	111	WD1	Tryptophan (W)	High	Moderate	Close to active site in Type 2 *P.en* and *P.nz*		High	Moderate		Yes	Good		Functionally important
	P. en, P. fs, P.nz , P.ch	155	WD2	Serine (S)	High	Moderate			High	Moderate		No	No		Functionally important
	P. en, P. fs, P.nz , P.ch	157	WD2	Aspartic acid (D)	High	Moderate			High	Moderate		Moderate	No		Functionally important
	P. en, P. fs, P.nz , P.ch	163	WD2	Tryptophan (W)	High	Moderate	Active site Type 2 *P.nz*		High	Moderate		Moderate	No		Functionally important
	P. en, P. fs, P.nz , P.ch	177	WD3	Histidine (H)	Moderate	High			High	Moderate		No	No		Functionally important
2	*P.ch*	190	SNP	Arginine (R) to Methionine (M)	Low	Low	Close to active site		Moderate	Moderate			No		R is more favourable and M is unfavourable per position
TTG1 Type 1 & Type 2	*P. en, P. fs, P.nz , P.ch*	198	WD4	Aspartic acid (D)	High	High			Moderate	Moderate		No	No		Functionally important
	P. en, P. fs, P.nz , P.ch	266	WD4	Histidine (H)	High	Moderate			High	Moderate	Allowed	No	No		Functionally important
	P. en, P. fs, P.nz , P.ch	277	WD4	Proline (P)	Moderate	Moderate	Close to active site in Type 2 *P.fs*		High	Moderate	Allowed	Yes	No		Functionally important
2	*P. en, P. fs*	284	WD4 SNP	Cysteine (C) to Serine (S)	Moderate	Low			High	High	Low	Moderate	No		C is more favourable and S is unfavourable per position
TTG1 Type 1 & Type 2	*P. en, P. fs, P.nz , P.ch*	289	WD4	Aspartic acid (D)	High	High	Close to active site in Type 2 *P.fs*		High	Moderate			No		Functionally important
	P. en, P. fs, P.nz , P.ch	295	WD4	Tryptophan (W)	High	Moderate	Close to active site in Type 2 *P.fs*		High	High		No	No		Functionally important
	P. en, P. fs, P.nz , P.ch	24	R2R3 MYB-SNP	Aspartic acid (D) to Asparagine (N)					High	High		No	No		D is the most favourable per position and N is very unfavourable
GL1 type 1 &	*P. en, P. fs, P.nz , P.ch*	66	R2R3 MYB-SNP	Asparagine (N) to Lycine (K)			Close to active site		High	Moderate		No	No		K is the most favourable per position and also close to active site
GL1 type 2	*P.nz*	109	R2R3 MYB-SNP	Tyrosine (Y) to Asparagine (N)					High			No	No		Y and H are the most favourable per position and N is unfavourable

3.4.4.2 GL1 proteins 3D model prediction and analysis of significant amino acid variations

GL1 models have shown high confidence and high sequence identity with the selected top template 'transcription factor WER' and are thus regarded as accurate according to Phyre2 (Table 3.3). The coverage with the template was comparatively low only covering half of the N-terminal R2R3 domain, and complete structure and functional analyses were not made (represented by the blank boxes in the table) in the program so adequate interpretations were hard to make. Therefore, more information gained from the literature is discussed below.

GO functional annotations of the top template and *A. thaliana* GL1 overlap, such as DNA binding and DNA binding transcription factor activity, even for the truncated *P. fastigiatum* Type 2 protein which ends in the middle of the N-terminal R2R3 domain. The amino acid alignment of the *A. thaliana* and *Pachycladon* species GL1 amino acid sequences are shown in Figure 3.8. The frameshift mutation caused by a five bp deletion in exon one of the *P. fastigiatum* GL1 Type 2 homeolog leads to 13 amino acid replacements and an early stop codon. Despite this, the Phyre2 in-depth structure and functional analysis predicted those are the most favorable residues adapted to survive the truncation and hence it may be still functional in a different aspect. Especially, the different amino acid residues added in the C-terminal end of *P. fastigiatum* are highly conserved, the last tryptophan (W) being the most conserved residue. Table 3.3 presents information on the unfavorable SNPs found in *Pachycladon* sequences concerning the *A. thaliana* sequence and the most conserved and favorable residues found in the sequences. The 24th positioned amino acid in Type 1 from negatively charged aspartic acid (D) (*A. thaliana*) to polar uncharged asparagine (N) and the 109th positioned SNP in Type 2 *P. novae-zealandiae* changed from uncharged hydrophobic tyrosine (Y) to Asparagine (N) were regarded as unfavorable. The last ten amino acids were found to be highly variable including a six-residual deletion in *P. enysii* Type 2 protein and histidine insertion in *P. enysii* and *P. fastigiatum* Type 1 proteins.

Table 3.3: The top protein 3D model predicted by Phyre2 for GL1 homeologs, PDB top template information with GO annotations and *A. thaliana* GO molecular function annotation

Homeolog	**Potential 3D model**	**Selected PDB top template information**							***A. thaliana* TTG1 GO molecular function annotation**
		PDB ID	**Confidence**	**Percent identity**	**PDB header**	**PDB molecule**	**PDB title**	**GO annotations**	
P. enysii GL1 Type 1	Proteins except P. fastigiatum GL1 Type 2  *P. fastigiatum* GL1 Type 2	6kks	100%	90%	DNA binding protein/DNA	transcription factor WER	structural insights into target dna recognition by r2r3-type myb2 transcription factor	nucleic acid binding, organic cyclic compound binding, heterocyclic compound binding, DNA-binding transcription factor activity, transcription regulator activity	DNA binding, DNA-binding transcription factor activity, protein binding, transcription cis-regulatory region binding
P. enysii GL1 Type 1			100%	92%					
P. fastigiatum GL1 Type 1			100%	90%					
P. fastigiatum GL1 Type 2			98.8%	70%					
P. novae-zealandiae GL1 Type 1			100%	89%					
P. novae-zealandiae GL1 Type 2			100%	90%					
P. cheesemanii GL1 Type 1			100%	90%					
P. cheesemanii GL1 Type 2			100%	92%					

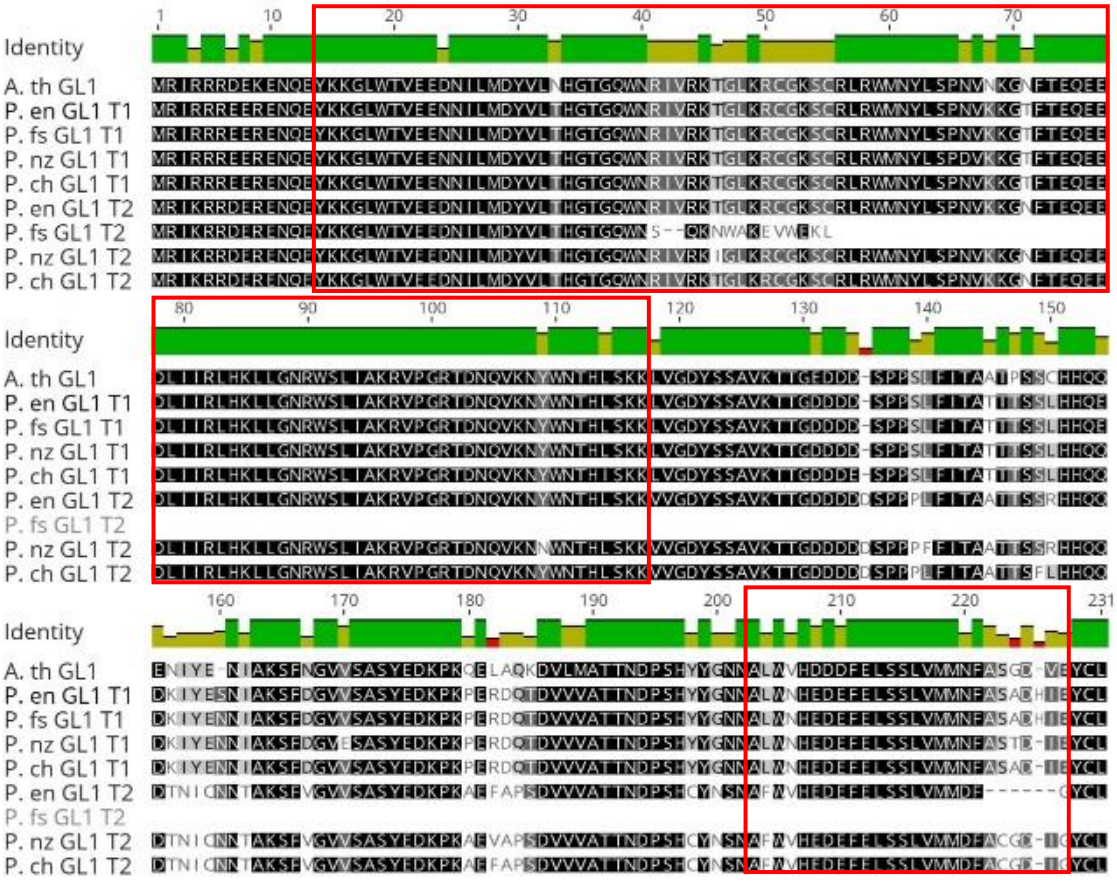


Figure 3.8: Alignment of *A. thaliana* and *Pachycladon* species GL1 amino acid sequences. Outlined in red are the functional motifs based on the *A. thaliana* sequence; residual positions 14-117 R2R3-MYB domain and 203-228 TAD. Positional information on domains was extracted from Uniprot (COUDERT et al. 2023) and previous studies (Bloomer et al 2012). The degree of sequence identity is shown on the top of the alignment. The species names are abbreviated here such as A. th - *A. thaliana*, P. ch - *P. cheesemanii*, P. en - *P. enysii*, P. nz - *P. novae-zealandiae* and P. fs - *P. fastigiatum*.

3.4.4.3 GL3 and EGL3 significant amino acid variations that could affect the functions of the proteins.

Although a proper protein functional analysis is not feasible with the incomplete homeologous sequence alignments of the two bHLH genes, some conspicuous mutations can be highlighted that could impact the functions of the proteins. We have complete protein copies of GL3 Type 1 all species, Type 2 *P. cheesemanii*, and EGL3 Type 1 and Type 2 *P. cheesemanii* to predict the effect of mutations.

From the sequences we have, it is obvious that the *P. cheesemanii GL3* Type 2 homeolog has a mutation right at the start codon ‘AAA’ instead of ‘ATG’, and a few frameshift

mutations. In addition, a 17 bp deletion at the start of exon 3 is observed in *P. enysii* and *P. fastigiatum* sequences, which would cause frameshift mutations, but the sequences we have been able to get are not long enough to identify an early stop codon. In contrast, *GL3* Type 1 sequences were found to be less variable with few polymorphisms observed in the *P. novae-zealandiae* and *P. cheesemanii* sequences.

The *EGL3* sequences we have been able to sequence do not have conspicuous indels and show only a few polymorphisms.

3.5 DISCUSSION

Pachycladon (Brassicaceae) has been identified as an allopolyploid lineage that formed ~1.6 mya (Joly et al 2009) from highly diverged parents, making this an ideal system for the examination of a duplicated genetic pathway. The trichome initiation pathway originally described from *A. thaliana* provides a good candidate system to explore the molecular evolution of a duplicated genetic pathway in *Pachycladon because* it has been well annotated in the close relative *A. thaliana*. Studying variation within and among homeologs of this pathway provides opportunities to reveal (1) just how diverged the pathways were prior to secondary contact and (2) how the two pathways have responded since that contact.

The conserved TTG1-R2R3 MYB-bHLH transactivation complex is the driving force of the trichome initiation pathway in *A. thaliana* and its relatives (Larkin et al 1994, Lin & Aoyama 2012, Pattanaik et al 2014, Payne et al 2000). The WD 40 protein TTG1, R2R3 MYB GL1, and bHLHs GL3 and EGL3 are the key proteins involved in the trichome initiation pathway. Here, the sequence variation in the two homeologs of each gene in four *Pachycladon* species was examined and compared to the model *A. thaliana*.

3.5.1 Gene polymorphisms create two distinct haplogroups

Haplotype networks exhibit the relationships between individual genotypes in a species or population (Leigh & Bryant 2015). The haplotype networks for *TTG1* and *GL1* were generated from the coding regions only and each yielded two haplogroups separated by many polymorphisms. For *GL1*, each homeolog from each species was distinguished by one to four polymorphisms, but for *TTG1*, the Type 1 sequences from *P. enysii* and *P. fastigiatum* were identical, reflecting the known close evolutionary relationship between these two species (Joly et al 2009). This same pattern was observed for each homeolog of each gene (Figure 3.4).

3.5.2 Type 2 gene copies have more variation than Type 1 gene copies

Based on the nucleotide diversity throughout the four functionally important WD domains of the *Pachycladon* TTG1 homeologs, it is not clear whether any of them are distinctly conserved. However, the C-terminal half of the WD3 domain and WD4 domain of both Types

are less variable compared to the other two WD domains suggesting the last two WD domains are functionally more critical for protein-protein interactions. Walker *et al* and Galway *et al* generated most of their glabrous and under-expressing trichome phenotypes by introducing SNPs in WD4 and the carboxyl end of *TTG1* (Galway et al 1994, Walker et al 1999). Overall variation suggests that the Type 1 homeolog is more conserved (nucleotide diversity 0.00648) than the Type 2 homeolog (nucleotide diversity 0.01053). Not many previous studies have focused on the nucleotide diversity or molecular evolution of TTG1, other than its interactions with GL3 and the functional importance of the domains (Payne et al 2000, Pesch et al 2015, Walker et al 1999, Zhao et al 2008), so comprehensive interpretations are hard to make.

GL1 Type 1 and Type 2 homeologs' estimates of nucleotide diversity from our samples are 0.009 and 0.013 are twice as high as that reported in one of the previous experiments (Pi = 0.0049) (Bloomer et al 2012) and more similar to that of another (Pi = 0.0109) study done (Hauser et al 2001), but both of the studies were done on *A. thaliana*. However, Bloomer *et al* (2012) analyzed *GL1* variation in 115 trichome-producing *A. thaliana* accessions, excluding glabrous accessions (in the sequence diversity analyses). But looking at the coding regions of our result, the pattern of the nucleotide diversity is quite similar to Bloomer et al (2012)'s observed pattern of low diversity within the domains and high in between them and in the 3' end, an expectation of any such analysis. Hauser et al (2001)'s value for nucleotide diversity is close to the results we obtained, but still, the study included a greater proportion of introns belonging to 28 *A. thaliana* accessions. We observed that the *GL1* Type 2 homeolog-wide value for nucleotide diversity is slightly larger than (0.013) that of the Type 1 homeolog-wide value (0.009), highlighting the presence of more polymorphisms in Type 2 homeologs. These observations support our hypothesis by demonstrating that *TTG1* and *GL1* Type 2 gene copies have more nucleotide diversity than Type 1 gene copies.

The nucleotide diversity of both *GL1* homeologs was quite low in the regions of the R2R3 MYB domain and TAD domain compared to the rest of the gene reflecting the

importance of domain structure and interaction stability for protein-protein and protein–DNA interactions. The mutations in the domains and their likely functional effects are discussed later.

3.5.3 The genes are under purifying selection

Ka⁄Ks sliding window comparisons between Type 1 and Type 2 *Pachycladon* homeologs of *TTG1* and *GL1* revealed that they are largely under purifying selection relative to each other across most of the coding region of the genes. However, the sliding window analyses revealed regions where Ka/Ks was greater than one in both *TTG1* and *GL1*, indicating certain regions are under positive selection, but necessarily do not belong in any functional motifs. Overall, the observed purifying selection on Type 2 copies aligns with our hypothesis of expected purifying selection on the Type 2 gene copies. However, the results suggest that Type 1 gene copies are also under purifying selection meaning that it helps to maintain the original function of the gene removing any mutations that could disrupt its role.

In *TTG1*, the sliding window analysis revealed that the coding region from 160-209 bp was under the influence of divergent selection. This result was underpinned by four non-synonymous mutations in the region (Type 1/AA position/Type 2): T178A, T182A, C202T and (GC)208C. Nayidu et al (2014) performed an evolutionary analysis to study the similarity and variation between five key trichome regulatory genes (*GL1*, *GL2*, *EGL3*, *TTG1* and *TRY*) of *B. villosa* and orthologues within three other *Brassicas* and *A. thaliana* using pairwise comparisons between the species of the coding regions. The Ka/Ks values for *TTG1* (including *EGL3*, *GL2* and *TRY*) were much less than one for most pair-wise comparisons between homologues and orthologues.

In *GL1*, the Ka/Ks spikes correspond with the regions from 427-482, 502-560, and 565-623. Interestingly, the first and last of these regions roughly overlap the results of Bloomer et al (2012) examining two diverged ‘Types’ of sequences in *A. thaliana*. The coding region corresponding to the first peak contains 10, the second peak 7 and the third peak 5 nonsynonymous mutations, all of which are located in exon 3. Screening the mutations that have the most impact will be discussed in detail in the protein analysis below. Ka/Ks ratios of

the *GL1* comparisons in *B. villosa* and each *Brassica* species and *Arabidopsis* of the above study of Nayidu et al (2014) were closer to one or more than one showing more synonymous substitutions and accelerated evolution indicating the potential for functional changes.

Tajima's test and Fu and Li's D* and F* test statistics indicated that the homeologs of *TTG1* and *GL1* are evolving neutrally, controversial to the purifying selection suggested by Ka/Ks analysis for the same genes. Small sample size, localized mutations, and limitations of the model assumptions could result in controversy. Similarly, Bloomer et al (2012)'s results on *GL1* molecular evolution did reject selective neutrality for *GL1* after Fu and Li's D* and F*, instead suggesting a role for purifying selection or population expansion. However, their dataset included a larger panel of *Arabidopsis thaliana* accessions (intraspecific variation) and non-coding regions, so direct comparisons are difficult to interpret.

3.5.4 Molecular evolution of the pathway proteins

3.5.4.1 Evolution of TTG1 duplicated protein copies

WD-40 repeats are important structural regions of TTG1 as they mediate protein-protein interactions (Walker et al 1999). However, none of the identified critical residues in WD repeats has been mutated in our dataset. Also, the C-terminal end of the protein is important for its proper functioning. The *ttg1-1* mutant, which is a truncated version of the protein lacking the C-terminal 25 amino acid residues, has very few trichomes on leaves and no anthocyanins in the testa or plant when homozygous (Galway et al 1994). This region is also conserved in our dataset. Several other truncated *ttg1* mutants have been formed with introduced termination codons for W-183 (*ttg1-19*), S-310 (*ttg1-15*, *ttg1-16*, *ttg1-17*, and *ttg1-18*) and S-310 together with S30C (*ttg1-20*) and also single base transition S282F (*ttg1-9*) (Walker et al 1999), all of which resulted in a *ttg1-1*-like phenotype. Based on modeling, only three amino acid substitutions P67A, R190M, and C284S found in the *Pachycladon* sequences were predicted to be unfavorable concerning function. Among them, only C284S was found to be included in a functional domain, which is WD4. The last two substitutions were found lying close to active sites so they may have some effects on protein structure and function.

TTG1 is a highly pleiotropic protein that is involved in multiple tasks such as positive regulation of trichome formation, anthocyanin production, seed coat mucilage production and seed coat pigmentation, and negative regulation of root hair formation (Pattanaik et al 2014, Ramsay & Glover 2005, Schiefelbein et al 2014, Xu et al 2015). Apart from these assigned functions, recently Airoldi et al (2019) have been shown that TTG1 is capable of partially regulating the circadian clock in the loss of function of other WD proteins in *Arabidopsis* LWD1 and LWD2. In addition, this does not seem to be a new function but an ancient role that was already there even before the duplication event created the WD paralogs. As such, supporting our hypothesis, that both of the TTG1 homeologs in the four *Pachycladon* species examined here appear to be functional and also under purifying selection.

3.5.4.2 Evolution of GL1 duplicated protein copies

Previous work by Stracke et al (2001) on the *A. thaliana* R2R3 MYB family deduced a consensus sequence using 125 *A. thaliana* R2R3-type MYB domain sequences, to identify the frequency of the most prevalent amino acids at each position to interpret the potential level of conservation and functional importance (Figure 3.9). The R2 and R3 regions contain three regularly spaced tryptophan residues forming a characteristic tryptophan cluster in each three-dimensional helix-turn-helix structure (Ogata et al 1992). There are five conserved tryptophan residues found altogether in the consensus sequence (phenylalanine replaces one of the six tryptophan residues), which are unmutated in the *Pachycladon* homeologs. *Pachycladon* Type 1 and Type 2 R2 MYB sequences all show the presence of these highly conserved residues except for the *P. fastigiatum* Type 2 protein, which terminates in the 42nd amino acid position. This truncated copy of *GL1* is missing approximately one-third of the R2 region and is completely missing the R3 domain. Other than the *P. fastigiatum* Type 2 copy, the R3 MYB consensus sequences from *Pachycladon* homeologs are well aligned with most residues in the *Arabidopsis* consensus sequence as shown in Figure 3.9. *Pachycladon* sequences also have the conserved phenylalanine replacing the first tryptophan in the R3 MYB domain. Based on the

occurrence of similar amino acid residues, the R3 domain is said to be slightly more conserved than the R2 domain (Stracke et al 2001).

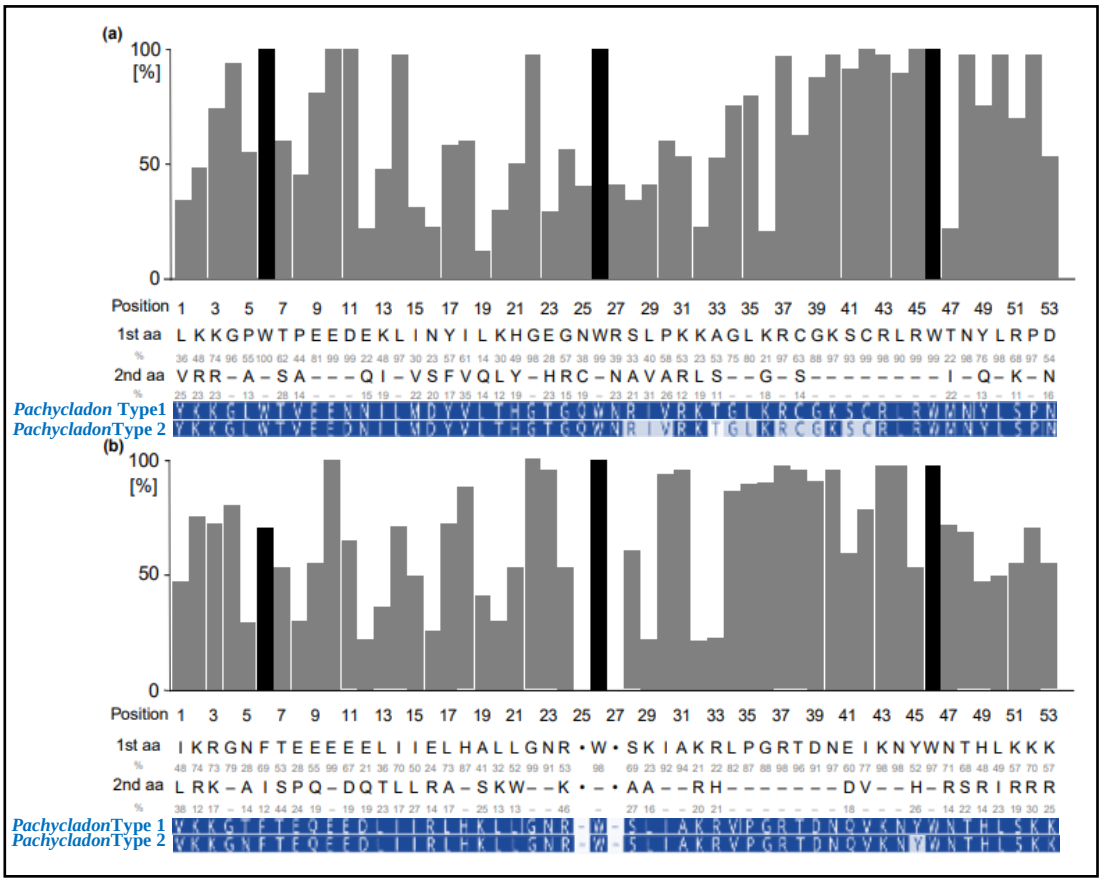


Figure 3.9: Consensus sequences and level of conservation of 125 *A. thaliana* R2 (a) and R3 (b) MYB domains extracted from Stracke et al (2001), aligned with *Pachycladon* GL1 Type 1 and Type 2 consensus sequences, shown in blue, derived from the protein sequences of *P. enysii*, *P. fastigiatum*, *P novae-zealandiae* and *P. cheesemanii*. The R3 MYB Type 2 *Pachycladon* consensus sequence does not include *P. fastigiatum* truncated sequence which forms a short protein until the R2 MYB domain. The *Arabidopsis* domains are aligned to show maximum identity by introducing two gaps, one towards amino- and the other towards carboxy-terminal from the second tryptophan (W) residue in the R3 MYB-repeat sequence. The percentage of the frequency of any amino acid found in each position of the alignment is indicated below the amino acid codes.

The GL1 C-terminal end histidine insertions in the *P. enysii* and *P. fastigiatum* Type 1 copies, multiple amino acid substitutions in Type 2 sequences and six amino acid deletions in *P. enysii* Type 2 are not expected to have functional effects because previous research has demonstrated that the last 14 amino acid residues of GL1 in C-terminal do not have a major

role on trichomes initiation (Hauser et al 2001). Further, the previously cloned *gl1-2* alleles which are lacking 27 amino acids in the C - terminal end caused only partial loss of protein activity which resulted in a reduction in leaf trichome initiation (Esch et al 1994), but this region is almost conserved in our dataset apart from the last 11 amino acids discussed previously. A negatively charged amino acid cluster 'DDD and E' in the 206-208 and 210th positions is regarded as an acidic activating region required for gene function (Esch et al 1994). This cluster is substituted in our dataset but with negatively charged residues, so a significant impact is not expected.

A conserved 20 amino acid signature in the R3 MYB domain from positions 78 to 98 has been identified as the structural motif important for the interaction between R2R3 MYBs and bHLH proteins, with special emphasis on the leucine (L) at position 86 and arginine (R) at position 101, as the two amino acids in the conserved motif that responsible for the most of the interaction strength (Zimmermann et al 2004). This region is conserved in our dataset except for the truncated *P. fastigiatum* Type 2 protein. Given its central role in trichome initiation, the discovery of this nonfunctional copy of *GL1* aligns with the glabrous phenotype of *P. fastigiatum*. Naturally occurring mutations in *GL1* that cause a glabrous phenotype have also been described in *A. thaliana* (Bloomer et al (2012) including a frameshift mutation. However, the other *GL1* homeolog from *P. fastigiatum* appears to be complete and would be expected to be functional so perhaps expression changes are also at play.

All Type 1 and Type 2 copies (obviously *P. fastigiatum* GL1 Type 2 must not be included) were predicted to have the same 3D model and the top template match from the protein database in Phyre2 analysis. The GO functional annotations were well overlapped with the GL1 paralog WER, which was predicted to be the top template. Hence, based on Phyre2 results, both copies of *P. enysii*, *P. novae-zealandiae,* and *P. cheesemanii* appear functionally redundant in the trichome initiation role. In addition, all of the copies were found to be under purifying selection.

3.5.4.3 Evolution of GL3 duplicated protein copies

GL3 has three distinct protein-protein interaction domains important in trichome initiation, the first 100 amino acids are required for the interaction with GL1, then the amino acids positioned in the range of 200-400 are needed for the interaction with TTG1, and the bHLH domain including the carboxy end fragment is required to interact with itself (homodimerize) (Payne et al 2000, Szymanski et al 2000). The trichome phenotype effects of *gl3* mutants are found to be less severe compared to *ttg1* and *gl1* mutants (Szymanski et al 2000) as trichomes are still produced but with fewer or no branches. The cause for this is believed to be the reduction of endoreduplication (Hülskamp et al 1994). The *gl3-1* mutant lacking a bHLH domain and nuclear localization signals (NLSs) motif, still could mediate interactions with GL1 and TTG1, suggesting that the bHLH domain is not responsible for the interactions with TTG1 and GL1 (Afrin et al 2014, Payne et al 2000, Pireyre & Burow 2015, Zhang et al 2003). The well-conserved bHLH domain is important for binding with DNA and homo- and heterodimeric protein-protein interactions (Atchley & Fitch 1997). The *gl3-2* allele produces a truncated polypeptide with only 46 amino acids. This mutant also yields less-branched trichomes (Payne et al 2000). The GL3 functions with regard to trichomes are somewhat redundant with EGL3, but when the double mutant is formed between *egl3-1* and *gl3-1*, a glabrous phenotype results (Zhang et al 2003).

P. enysii, *P. fastigiatum* and *P. cheesemanii* and potentially *P. novae-zealandiae* (but sequences are incomplete) *GL3* Type 2 sequences in our data have some large deletions as mentioned in the results above in the range of 1-400 bp region which could have an impact on GL1 and TTG1 binding. Also, the carboxyl end of all Type 2 copies may not be available if they are short proteins ending at 413 amino acids long as suggested by our sequences, which would affect homodimerization. Significant SNPs involving A108P in GL3 Type 2 *P. enysii*, *P. fastigiatum* and *P. nove-zealandiae,* and A109S in all GL3 Type 2 homeologs since proline and serine substitutions can cause potential conformational changes.

3.5.4.4 Evolution of EGL3 duplicated protein copies

EGL3 encodes a protein length of 596 amino acids, which is only 39 amino acids shorter than GL3, and shows nearly 75% similarity to GL3 at the amino acid level (Zhang et al 2003). Also, the *egl3* single mutant produces a subtle trichome phenotype not easily noticeable when GL3 is present (Zhang et al 2003). The first 367 amino acids are required to interact with TTG1 and MYB domains of GL1, and the last 229 amino acids in the carboxy fragment interact with itself and bHLH end of GL3 (Zhang et al 2003). However, no significant SNPs or indels are found in the *Pachycladon EGL3* sequence(s).

When the gene constituents in genetic pathways are duplicated after WGD, according to the gene balance hypothesis, functional modules would co-evolve along with the other regulators to maintain a balanced relationship through the co-evolution of protein-protein interactions (Lemos et al 2004).

3.5.4.5 Genetic redundancy of duplicates and functional divergence

Genetic redundancy is the immediate outcome of gene duplication, but over time, one of the duplicates may be lost or both duplicates may be retained and evolve to subfunctionalize or neofunctionalize or be maintained when the dosage effect presents a selective advantage (Briggs et al 2006, Chen et al 2010, Osborn et al 2003, Pérez-Pérez et al 2009, Pickett & Meeks-Wagner 1995). According to the gene dosage balance model, the genes encoding the interacting proteins in transcriptional regulation, signal transduction, and complexes are predicted to be co-retained after WGD and to experience purifying selection leading to prolonged retention (Edger & Pires 2009, Veitia et al 2013). First, this can be executed with only one copy as in Figure 3.10 which continues to affect the whole complex to develop a new favorable balance later (Birchler et al 2007).

Considering our dataset, crucial disruption in the structure was seen only in the *P. fastigiatum* GL1 Type 2 protein while both copies of *TTG1* seem to be fully retained. At this stage, it is unclear whether the truncated *GL1* homeolog found in *P. fastigiatum* is sufficient to disrupt trichome initiation by itself, potentially due to a gene dosage issue, or if the other

homeolog, which has no obviously detrimental amino acid replacements, also has functional issues; e.g., with its expression. For the genes for which we have a complete data set, *GL1*, and *TTG1*, the Type 2 homeologs have greater variation than the Type 1 homeologs. Whether this reflects some relaxed selection pressure due to duplication is unclear. The potential differential expression of homeologs in the pathway is discussed in Chapter 4.

Figure 3.10: Model showing the effect of the stoichiometry of gene regulatory complexes and its impact in generating phenotypes reproduced from Birchler et al (2007). Panel A shows a stoichiometrically balanced regulatory complex shown by sets of rectangles, circles, and triangles (XX: YY: ZZ::1:1:1), which interacts with the *cis* factor of a target structural gene, that results in the normal phenotype shown by a thick green rectangle. In panel B the stoichiometry of the regulatory complex is altered and unbalanced (XX: Y: ZZ::2:1:2) so the interaction with the *cis*-factor of the target loci is impacted that results in an abnormal phenotype, which is indicated by a thinner green rectangle. This phenotype might be selected against in the purifying selection. In panel C, the *cis* factor is mutated (depicted by a brown rectangle) and the unbalanced status of the regulatory complex is tolerated to recover the normal phenotype. Eventually, the unbalanced regulatory complex is co-adapted to suit the mutation of the target gene and gains a new balance (X: Y: Z::1:1:1) to maintain the normal phenotype as shown in panel D.

3.5 CONCLUSION

The allopolyploid *Pachycladon* possesses highly diverged duplicated copies of *TTG1*, *GL1*, *GL3*, and *EGL3*, which participate in forming a trichome initiation complex. The evolution of the duplicates is found to be associated with functional attributes including pleiotropy of individual proteins such as TTG1 and epistasis among all interacting proteins. Both copies of the highly pleiotropic *TTG1* tend to be retained compared to the less pleiotropic *GL1* in the young allopolyploid, at least in the present stage of evolution. The Type 2 copy of GL1 in *Pachycladon fastigiatum* is truncated, which would disable interactions with other proteins to build the transactivation complex. Likewise, one copy of all the proteins seems to be accumulating more mutations than the other, namely the more diverged Type 2 copies. It can be assumed the complex and the pathway is undergoing coordinated evolution which will drive a new balance among the proteins. In other words, the alternative copies may be eliminated from the trichome initiation complex through fractionation or maintained through functional divergence such as subfunctionalization or neofunctionalization to be involved in different roles in different pathways.

REFERENCES

Afrin, S., M. Nuruzzaman, J. Zhu and Z.-Y. Luo, 2014 Combinatorial interactions of MYB and bHLH in flavonoid biosynthesis and their function in plants, pp.

Airoldi, C. A., T. J. Hearn, S. F. Brockington, A. A. R. Webb and B. J. Glover, 2019 TTG1 proteins regulate circadian activity as well as epidermal cell fate and pigmentation. Nat Plants 5**:** 1145-1153.

Atchley, W. R., and W. M. Fitch, 1997 A natural classification of the basic helix-loop-helix class of transcription factors. Proceedings of the National Academy of Sciences 94**:** 5172-5176.

Bae, H., and J. Coller, 2022 Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. Molecular cell 82**:** 1467-1476.

Birchler, J. A., H. Yao and S. Chudalayandi, 2007 Biological consequences of dosage dependent gene regulatory systems. Biochim Biophys Acta 1769**:** 422-428.

Blanc, G., and K. H. Wolfe, 2004 Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16**:** 1679-1691.

Bloomer, R. H., T. E. Juenger and V. V. Symonds, 2012 Natural variation in GL1 and its effects on trichome density in Arabidopsis thaliana. Molecular Ecology 21**:** 3501-3515.

Bowman, J. L., L. N. Briginshaw, T. J. Fisher and E. Flores-Sandoval, 2019 Something ancient and something neofunctionalized—evolution of land plant hormone signaling pathways. Current Opinion in Plant Biology 47**:** 64-72.

Briggs, G. C., K. S. Osmont, C. Shindo, R. Sibout and C. S. Hardtke, 2006 Unequal genetic redundancies in Arabidopsis--a neglected phenomenon? Trends Plant Sci 11**:** 492-498.

Capra, J. A., and M. Singh, 2007 Predicting functionally important residues from sequence conservation. Bioinformatics 23**:** 1875-1882.

Carretero-Paulet, L., and M. A. Fares, 2012 Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Molecular biology and evolution 29**:** 3541-3551.

Chapman, B. A., J. E. Bowers, F. A. Feltus and A. H. Paterson, 2006 Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proceedings of the National Academy of Sciences 103**:** 2730-2735.

Chen, H.-W., S. Bandyopadhyay, D. E. Shasha and K. D. Birnbaum, 2010 Predicting genome-wide redundancy using machine learning. BMC Evolutionary Biology 10**:** 357.

Cheng, Z., L. Sun, T. Qi, B. Zhang, W. Peng *et al.*, 2011 The bHLH transcription factor MYC3 interacts with the Jasmonate ZIM-domain proteins to mediate jasmonate response in Arabidopsis. Mol Plant 4**:** 279-288.

Clark, J. W., and P. C. Donoghue, 2018 Whole-genome duplication and plant macroevolution. Trends in plant science 23**:** 933-945.

Conant, G. C., and K. H. Wolfe, 2008 Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics 9**:** 938-950.

Coudert, E., S. Gehant, E. de Castro, M. Pozzato, D. Baratin *et al.*, 2023 Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics 39.

Dayhoff, M. O., and R. V. Eck, 1972 *Atlas of protein sequence and structure*. National Biomedical Research Foundation.

De Smet, R., E. Sabaghian, Z. Li, Y. Saeys and Y. Van de Peer, 2017 Coordinated Functional Divergence of Genes after Genome Duplication in Arabidopsis thaliana The Plant Cell 29**:** 2786-2800.

Del Pozo, J. C., and E. Ramirez-Parra, 2015 Whole genome duplications in plants: an overview from Arabidopsis. Journal of Experimental Botany 66**:** 6991-7003.

Dombrecht, B., G. P. Xue, S. J. Sprague, J. A. Kirkegaard, J. J. Ross *et al.*, 2007 MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. Plant Cell 19**:** 2225-2245.

Dubos, C., R. Stracke, E. Grotewold, B. Weisshaar, C. Martin *et al.*, 2010 MYB transcription factors in Arabidopsis. Trends in plant science 15**:** 573-581.

Edger, P. P., and J. C. Pires, 2009 Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 17**:** 699-717.

Esch, J. J., D. G. Oppenheimer and M. D. Marks, 1994 Characterization of a weak allele of the GL1 gene of Arabidopsis thaliana. Plant Molecular Biology 24**:** 203-207.
Fay, J. C., and C. I. Wu, 2001 The neutral theory in the genomic era. Curr Opin Genet Dev 11**:** 642-646.
Foxe, J. P., V. U. Dar, H. Zheng, M. Nordborg, B. S. Gaut *et al.*, 2008 Selection on amino acid substitutions in Arabidopsis. Mol Biol Evol 25**:** 1375-1383.
Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133**:** 693-709.
Galway, M. E., J. D. Masucci, A. M. Lloyd, V. Walbot, R. W. Davis *et al.*, 1994 The TTG gene is required to specify epidermal cell fate and cell patterning in the Arabidopsis root. Developmental biology 166**:** 740-754.
Geneious 9.1.8, 2017, pp.
Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. science 185**:** 862-864.
Gray, V. E., R. J. Hause and D. M. Fowler, 2017 Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. Genetics 207**:** 53-61.
Guo, H., T.-H. Lee, X. Wang and A. H. Paterson, 2013 Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiology 162**:** 769-778.
Han, G., Y. Li, Z. Yang, C. Wang, Y. Zhang *et al.*, 2022 Molecular mechanisms of plant trichome development. Frontiers in Plant Science 13**:** 910228.
Hauser, M.-T., B. Harr and C. Schlötterer, 2001 Trichome distribution in Arabidopsis thaliana and its close relative Arabidopsis lyrata: molecular analysis of the candidate gene GLABROUS1. Molecular Biology and Evolution 18**:** 1754-1763.
Hendriks, K. P., C. Kiefer, I. A. Al-Shehbaz, C. D. Bailey, A. H. van Huysduynen *et al.*, 2023 Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. Current Biology.
Henikoff, S., and J. G. Henikoff, 1992 Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89**:** 10915-10919.
Huang, C.-H., R. Sun, Y. Hu, L. Zeng, N. Zhang *et al.*, 2015 Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. Molecular Biology and Evolution 33**:** 394-412.
Hülskamp, M., S. Misŕa and G. Jürgens, 1994 Genetic dissection of trichome cell development in Arabidopsis. Cell 76**:** 555-566.
Hurst, L. D., 2002 The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 18**:** 486.
Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. Nature 473**:** 97.
Joly, S., P. B. Heenan and P. J. Lockhart, 2009 A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. Mol Phylogenet Evol 51**:** 365-372.
Jones, D. M., and K. Vandepoele, 2020 Identification and evolution of gene regulatory networks: insights from comparative studies in plants. Current opinion in plant biology 54**:** 42-48.
Kaltenegger, E., S. Leng and A. Heyl, 2018 The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. BMC Evolutionary Biology 18**:** 76.
Kelley, L. A., S. Mezulis, C. M. Yates, M. N. Wass and M. J. E. Sternberg, 2015 The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10**:** 845-858.
Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.
Kirik, V., M. M. Lee, K. Wester, U. Herrmann, Z. Zheng *et al.*, 2005 Functional diversification of MYB23 and GL1 genes in trichome morphogenesis and initiation. Development 132**:** 1749-1749.
Kirik, V., A. Schnittger, V. Radchuk, K. Adler, M. Hülskamp *et al.*, 2001 Ectopic expression of the Arabidopsis AtMYB23 gene induces differentiation of trichome cells. Dev Biol 235**:** 366-377.
Larkin, J. C., D. G. Oppenheimer, A. M. Lloyd, E. T. Paparozzi and M. D. Marks, 1994 Roles of the GLABROUS1 and TRANSPARENT TESTA GLABRA Genes in Arabidopsis Trichome Development. The Plant cell 6**:** 1065-1076.
Larkin, J. C., J. D. Walker, A. C. Bolognesi-Winfield, J. C. Gray and A. R. Walker, 1999 Allele-specific interactions between ttg and gl1 during trichome development in Arabidopsis thaliana. Genetics 151**:** 1591-1604.

Lee, M. M., and J. Schiefelbein, 1999 WEREWOLF, a MYB-related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning. Cell 99**:** 473-483.
Lee, M. M., and J. Schiefelbein, 2001 Developmentally distinct MYB genes encode functionally equivalent proteins in Arabidopsis. Development 128**:** 1539-1546.
Leigh, J. W., and D. Bryant, 2015 popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 6**:** 1110-1116.
Lemos, B., C. D. Meiklejohn and D. L. Hartl, 2004 Regulatory evolution across the protein interaction network. Nat Genet 36**:** 1059-1060.
Li, J., Z. Zhang, S. Vang, J. Yu, G. K.-S. Wong *et al.*, 2009 Correlation Between Ka/Ks and Ks is Related to Substitution Model and Evolutionary Lineage. Journal of Molecular Evolution 68**:** 414-423.
Li, L., R. Briskine, R. Schaefer, P. S. Schnable, C. L. Myers *et al.*, 2016 Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. BMC genomics 17**:** 1-15.
Li, W.-H., J. Yang and X. Gu, 2005 Expression divergence between duplicate genes. TRENDS in Genetics 21**:** 602-607.
Liang, Z., and J. C. Schnable, 2018 Functional divergence between subgenomes and gene pairs after whole genome duplications. Molecular plant 11**:** 388-397.
Lin, Q., and T. Aoyama, 2012 Pathways for epidermal cell differentiation via the homeobox gene GLABRA2: update on the roles of the classic regulator. J Integr Plant Biol 54**:** 729-737.
Magadum, S., U. Banerjee, P. Murugan, D. Gangapur and R. Ravikesavan, 2013 Gene duplication as a major force in evolution. Journal of Genetics 92**:** 155-161.
Makałowski, W., and M. S. Boguski, 1998 Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proceedings of the National Academy of Sciences 95**:** 9407-9412.
Mandakova, T., S. Joly, M. Krzywinski, K. Mummenhoff and M. A. Lysak, 2010 Fast diploidization in close mesopolyploid relatives of Arabidopsis. Plant Cell 22**:** 2277-2290.
McGrath, C., and M. Lynch, 2012 Evolutionary significance of whole-genome duplication, pp. 1-20 in *Polyploidy and genome evolution*. Springer.
Mershon, J. P., M. Becker and C. P. Bickford, 2015 Linkage between trichome morphology and leaf optical properties in New Zealand alpine Pachycladon (Brassicaceae). New Zealand Journal of Botany 53**:** 175-182.
Moghe, G. D., D. E. Hufnagel, H. Tang, Y. Xiao, I. Dworkin *et al.*, 2014 Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish Raphanus raphanistrum and three other Brassicaceae species. The Plant Cell 26**:** 1925-1937.
Morohashi, K., M. Zhao, M. Yang, B. Read, A. Lloyd *et al.*, 2007 Participation of the Arabidopsis bHLH Factor GL3 in Trichome Initiation Regulatory Events. Plant Physiology 145**:** 736-746.
Nayidu, N. K., S. Kagale, A. Taheri, T. S. Withana-Gamage, I. A. Parkin *et al.*, 2014 Comparison of five major trichome regulatory genes in Brassica villosa with orthologues within the Brassicaceae. PLoS One 9**:** e95877.
Nekrutenko, A., K. D. Makova and W. H. Li, 2002 The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res 12**:** 198-202.
Nikolov, L. A., P. Shushkov, B. Nevado, X. Gan, I. A. Al-Shehbaz *et al.*, 2019 Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytologist 222**:** 1638-1651.
Nougué, O., J. Corbi, S. G. Ball, D. Manicacci and M. I. Tenaillon, 2014 Molecular evolution accompanying functional divergence of duplicated genes along the plant starch biosynthesis pathway. BMC evolutionary biology 14**:** 1-16.
Ogata, K., H. Hojo, S. Aimoto, T. Nakai, H. Nakamura *et al.*, 1992 Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. Proceedings of the National Academy of Sciences 89**:** 6428-6432.
Oppenheimer, D. G., P. L. Herman, S. Sivakumaran, J. Esch and M. D. Marks, 1991 A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. Cell 67**:** 483-493.

Osborn, T. C., J. C. Pires, J. A. Birchler, D. L. Auger, Z. J. Chen *et al.*, 2003 Understanding mechanisms of novel gene expression in polyploids. Trends in genetics 19**:** 141-147.
Paffendorf, B. A., R. Qassrawi, A. M. Meys, L. Trimborn and A. Schrader, 2020 TRANSPARENT TESTA GLABRA 1 participates in flowering time regulation in Arabidopsis thaliana. PeerJ 8**:** e8303.
Patra, B., S. Pattanaik and L. Yuan, 2013 Ubiquitin protein ligase 3 mediates the proteasomal degradation of GLABROUS 3 and ENHANCER OF GLABROUS 3, regulators of trichome development and flavonoid biosynthesis in Arabidopsis. The Plant Journal 74**:** 435-447.
Pattanaik, S., B. Patra, S. K. Singh and L. Yuan, 2014 An overview of the gene regulatory network controlling trichome development in the model plant, Arabidopsis. Frontiers in plant science 5**:** 259.
Payne, C. T., F. Zhang and A. M. Lloyd, 2000 GL3 Encodes a bHLH Protein That Regulates Trichome Development in Arabidopsis Through Interaction With GL1 and TTG1. Genetics 156**:** 1349-1362.
Pérez-Pérez, J. M., H. Candela and J. L. Micol, 2009 Understanding synergy in genetic interactions. Trends Genet 25**:** 368-376.
Pesch, M., and M. Hülskamp, 2009 One, two, three…models for trichome patterning in Arabidopsis? Current Opinion in Plant Biology 12**:** 587-592.
Pesch, M., I. Schultheiß, K. Klopffleisch, J. F. Uhrig, M. Koegl *et al.*, 2015 TRANSPARENT TESTA GLABRA1 and GLABRA1 Compete for Binding to GLABRA3 in Arabidopsis. Plant Physiology 168**:** 584-597.
Pickett, F. B., and D. R. Meeks-Wagner, 1995 Seeing double: appreciating genetic redundancy. The Plant Cell 7**:** 1347-1356.
Pireyre, M., and M. Burow, 2015 Regulation of MYB and bHLH Transcription Factors: A Glance at the Protein Level. Molecular Plant 8**:** 378-388.
Presnyak, V., N. Alhusaini, Y.-H. Chen, S. Martin, N. Morris *et al.*, 2015 Codon optimality is a major determinant of mRNA stability. Cell 160**:** 1111-1124.
Qiao, X., Q. Li, H. Yin, K. Qi, L. Li *et al.*, 2019 Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome biology 20**:** 1-23.
Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan, 1963 Stereochemistry of polypeptide chain configurations. J Mol Biol 7**:** 95-99.
Ramsay, N. A., and B. J. Glover, 2005 MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. Trends in Plant Science 10**:** 63-70.
Ray, A., E. Lindahl and B. Wallner, 2012 Improved model quality assessment using ProQ2. BMC Bioinformatics 13**:** 224.
Riechmann, J. L., J. Heard, G. Martin, L. Reuber, C.-Z. Jiang *et al.*, 2000 Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. science 290**:** 2105-2110.
Romero, I., A. Fuertes, M. Benito, J. Malpica, A. Leyva *et al.*, 1998 More than 80R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. The Plant journal: for cell and molecular biology 14**:** 273-284.
Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light *et al.*, 2007 Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 308**:** 58-73.
Roulin, A., P. L. Auer, M. Libault, J. Schlueter, A. Farmer *et al.*, 2013 The fate of duplicated genes in a polyploid plant genome. The Plant Journal 73**:** 143-153.
Rozas, J., A. Ferrer-Mata, J. C. Sánchez-DelBarrio, S. Guirao-Rico, P. Librado *et al.*, 2017 DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Molecular Biology and Evolution 34**:** 3299-3302.
Sankoff, D., and C. Zheng, 2018 Whole genome duplication in plants: implications for evolutionary analysis. Comparative Genomics: Methods and Protocols**:** 291-315.
Schiefelbein, J., X. Zheng and L. Huang, 2014 Regulation of epidermal cell fate in Arabidopsis roots: the importance of multiple feedback loops. Frontiers in Plant Science 5.
Schiessl, S., 2020 Regulation and Subfunctionalization of Flowering Time Genes in the Allotetraploid Oil Crop Brassica napus. Frontiers in Plant Science 11.

Schmidtke, P., V. Le Guilloux, J. Maupetit and P. Tuffï¿ ½ry, 2010 Fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic acids research 38**:** W582-W589.
Schnittger, A., U. Folkers, B. Schwab, G. Jürgens and M. Hülskamp, 1999 Generation of a Spacing Pattern: The Role of TRIPTYCHON in Trichome Patterning in Arabidopsis. The Plant Cell 11**:** 1105-1116.
Sémon, M., and K. H. Wolfe, 2007 Consequences of genome duplication. Current opinion in genetics & development 17**:** 505-512.
Signor, S. A., and S. V. Nuzhdin, 2018 The Evolution of Gene Expression in cis and trans. Trends in Genetics 34**:** 532-544.
Soltis, P. S., and D. E. Soltis, 2016 Ancient WGD events as drivers of key innovations in angiosperms. Current opinion in plant biology 30**:** 159-165.
Srinivasan, N., R. Sowdhamini, C. Ramakrishnan and P. Balaram, 1990 Conformations of disulfide bridges in proteins. International Journal of Peptide and Protein Research 36**:** 147-155.
Stracke, R., M. Werber and B. Weisshaar, 2001 The R2R3-MYB gene family in Arabidopsis thaliana. Current Opinion in Plant Biology 4**:** 447-456.
Szymanski, D. B., A. M. Lloyd and M. D. Marks, 2000 Progress in the molecular genetic analysis of trichome initiation and morphogenesis in Arabidopsis. Trends Plant Sci 5**:** 214-219.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585-595.
Telfer, A., K. M. Bollman and R. S. Poethig, 1997 Phase change and the regulation of trichome distribution in Arabidopsis thaliana. Development 124**:** 645-654.
Van de Peer, Y., S. Maere and A. Meyer, 2009 The evolutionary significance of ancient genome duplications. Nat Rev Genet 10**:** 725-732.
Veitia, R. A., S. Bottani and J. A. Birchler, 2013 Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. Trends in Genetics 29**:** 385-393.
Walker, A. R., P. A. Davison, A. C. Bolognesi-Winfield, C. M. James, N. Srinivasan *et al.*, 1999 The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. The Plant cell 11**:** 1337-1350.
Wang, S., and J.-G. Chen, 2008 Arabidopsis Transient Expression Analysis Reveals that Activation of GLABRA2 May Require Concurrent Binding of GLABRA1 and GLABRA3 to the Promoter of GLABRA2. Plant and Cell Physiology 49**:** 1792-1804.
Wang, Y., S. P. Ficklin, X. Wang, F. A. Feltus and A. H. Paterson, 2016 Large-scale gene relocations following an ancient genome triplication associated with the diversification of core eudicots. PLoS One 11**:** e0155637.
Wang, Y., Q. Zhou, Z. Meng, M. A. Abid, Y. Wang *et al.*, 2022 Multi-Dimensional Molecular Regulation of Trichome Development in Arabidopsis and Cotton. Frontiers in Plant Science 13**:** 892381.
Ward, J. J., L. J. McGuffin, K. Bryson, B. F. Buxton and D. T. Jones, 2004 The DISOPRED server for the prediction of protein disorder. Bioinformatics 20**:** 2138-2139.
Xu, W., C. Dubos and L. Lepiniec, 2015 Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. Trends Plant Sci 20**:** 176-185.
Yates, C. M., I. Filippis, L. A. Kelley and M. J. E. Sternberg, 2014 SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. Journal of Molecular Biology 426**:** 2692-2701.
Yogeeswaran, K., C. Voelckel, S. Joly and P. B. Heenan, 2011 Pachycladon. 227-249.
Yu, Z., C. Zheng, V. A. Albert and D. Sankoff, 2020 Excision dominates pseudogenization during fractionation after whole genome duplication and in gene loss after speciation in plants. Frontiers in Genetics 11**:** 603056.
Zhang, F., A. Gonzalez, M. Zhao, C. T. Payne and A. Lloyd, 2003 A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. Development 130**:** 4859-4869.
Zhang, Z., J. Li, X.-Q. Zhao, J. Wang, G. K.-S. Wong *et al.*, 2006 KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. Genomics, Proteomics & Bioinformatics 4**:** 259-263.

Zhao, L., L. Gao, H. Wang, X. Chen, Y. Wang *et al.*, 2013 The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis. Funct Integr Genomics 13**:** 75-98.

Zhao, M., K. Morohashi, G. Hatlestad, E. Grotewold and A. Lloyd, 2008 The TTG1-bHLH-MYB complex controls trichome cell fate and patterning through direct targeting of regulatory loci. Development 135**:** 1991-1999.

Zimmermann, I. M., M. A. Heim, B. Weisshaar and J. F. Uhrig, 2004 Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. Plant J 40**:** 22-34.

A3. APPENDIX

A3.1 Statement of contribution

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Thilini Madara Warusawithana
Name and title of main supervisor:	Dr. V. Vaughan Symonds
In which chapter is the manuscript/published work?	Chapter 3
What percentage of the manuscript/published work was contributed by the student?	85%

Describe the contribution that the student has made to the manuscript/published work:

Thilini worked with Dr. Vaughan Symonds to develop the research question and objectives, and to design the experimental setup and data collection methods. Thilini conducted a comprehensive literature review to understand previous different analysis methods. Under the guidance of Dr. Symonds, Thilini also conducted analyses and interpreted the results. Thilini drafted the chapter and incorporated the feedback from Dr. Symonds and co-supervisor Dr. Matthieu Vignes.

Please select one of the following three options:

- ○ **The manuscript/published work is published or in press**
  Please provide the full reference of the research output:
- ○ **The manuscript is currently under review for publication**
  Please provide the name of the journal:
- ◉ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Thilini Warusawithana Digitally signed by Thilini Warusawithana Date: 2024.10.25 15:02:46 +13'00'	Main supervisor's signature:	Vaughan Symonds Digitally signed by Vaughan Symonds Date: 2024.10.29 11:43:39 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

1 Doctoral Research Committee July 2022

4.0 CHAPTER FOUR

Expression analysis of duplicated gene copies (homeologs) in *Pachycladon* trichome initiation pathway

4.1 ABSTRACT

When highly diverged lineages form allopolyploid lineages, the new genome undergoes many structural and regulatory changes that impact the composition of the genome and, by extension, of genetic pathways and the expression profiles of pathway components. As there has been relatively little work focused specifically on duplicated pathways, here we attempt to explore this phenomenon in the *Pachycladon* trichome initiation pathway by examining the expression profiles of the two homeologs of four genes of interest that play a central role in trichome initiation. qPCR analyses were completed for the homeologs of *TTG1*, *GL1*, *GL3*, and *EGL3* in *P. enysii*, *P. fastigiatum*, *P. novae-zealandiae*, and *P. cheesemanii*, using whole seedlings from two species and mature leaves of three species (including both stages from *P. cheesemanii*). The results suggest multiple molecular mechanisms overlap to control the expression of homeologs such as relative gene dosage balance, subgenome dominance, and pleiotropy. *GL3* and *GL1* showed biased expression of the Type 1 homeolog, especially *GL3*, which significantly overexpressed the Type 1 copy in all species in both seedlings and mature leaves. Available evidence suggests that the Type 2 copies are maternal in origin, indicating that the strongest pattern observed is one of paternal expression bias. However, the highly pleiotropic gene, *TTG1*, showed little to no evidence of biased expression for either Type. In the only species for which both seedlings and mature leaves were examined, *P. cheesemanii*, the expression of both homeologs of *GL1* was much greater in seedlings than in mature leaves. This matches findings in *A. thaliana* and the known trichome initiation role for *GL1*. The results reported here provide valuable insights into the molecular evolution of a genetic pathway following duplication.

4.2 INTRODUCTION

4.2.1 Polyploidy, differential expression and evolution

Polyploidy or whole genome duplication doubles the genome content relative to the progenitor lineage (Segraves 2017, Sémon & Wolfe 2007). In the case of a diploid progenitor or progenitors, this results in cells having more than two complete copies of all chromosomes (Negrón-Ortiz 2007, Sattler et al 2016). The redundancy of genetic material indicates a potential evolutionary advantage conferred to the polyploid lineage through subsequent processes of adaptive evolution by changing the genome structure, gene expression, and gene function that ultimately alter phenotypes (Liu et al 2017, Van de Peer et al 2017).Those changes such as increases in biomass, stature and growth rate are more easily recognized in allopolyploids compared to autopolyploids (Chen 2007, Chen 2010). One of the examples of adaptive evolution is the differential expression of homeologous genes in genetic pathways or regulatory networks (Kliebenstein 2008, Nougué et al 2014, VanderSluis et al 2010).

While allopolyploidy can occur via several routes, both hybridization and genome doubling are expected to lead to transcriptomic changes (Chen 2007, Flagel et al 2008, Wang et al 2006). Genetic and epigenetic changes may occur immediately after interspecific hybridization or allopolyploidization, including chromatin modifications through RNA interference and dosage dependency (additive effect due to increased gene content) leading to gene expression changes (Birchler et al 2001, Chen 2007).

When allopolyploids undergo these events, one or both duplicated gene copies are retained (Lynch & Conery 2000). One copy may become lost by the accumulation of mutations over time (pseudogenization) or both copies may be retained (Muller 1925). Gene expression patterns are believed to change faster during adaptation in response to external abiotic and biotic stresses than to internal developmental programs such as internal signals and developmental switches (Ha et al 2007). Subsequently, expression divergence between gene duplicates may facilitate functional divergence such as subfunctionalization and neofunctionalization (Lynch & Force 2000, Lynch et al 2001, Panchy et al 2016) and each of

the gene duplicates can become prominent in different tissues and different developmental stages (Madlung 2013).

4.2.2 Evolution of expression divergence in genetic complexes/pathways following whole genome duplication

4.2.2.1 Gene dosage balance hypothesis

When an expression change occurs in one of the interacting partners or target genes involved in a genetic network, that can result in an expression change in the rest of the components of the network, because of the interdependent activity of the components based on their relative gene dosage balance. This is referred to as the relative gene dosage balance hypothesis. (De Smet & Van de Peer 2012). When the homeologs are involved in crucial genetic complexes or pathways, the expression divergence rate could be relatively slow enabling the complex to fine-tune according to dosage effect, making it resistant to detrimental effects (Bomblies & Doebley 2006, Chen 2007, Liang & Schnable 2018, Thomas et al 2006).

One way that genetic pathways evolve following duplication is that the mutations occur in one homeologous member of the complex affecting interacting partners, so the selection pressure is gradually added on the complex, creating subtle variations on other balanced interactors to shift the stoichiometric relationships into a new balance. This has been described in some experiments (Levine et al 2007, Rodriguez et al 2007) conducted on male-specific lethal complex in *Drosophila*. Another molecular evolutionary route is cis-regulatory or transcription target site divergence of downstream genes of the complex that in turn would influence the complex to evolve (Birchler et al 2007, Lemos et al 2008). MicroRNAs could participate in modulating the gene expression as required in these scenarios (Wang & Adams 2015). The constraints imposed on duplicated regulatory genes via these mechanisms can hold them in an evolutionary lineage for a longer period than other classes of genes (Birchler & Veitia 2010). Many species show evidence of the gene balance hypothesis by exhibiting significantly correlated expression patterns such as in *Arabidopsis* (Blanc & Wolfe 2004a), rice

(Throude et al 2009, Yim et al 2009), and cotton (Renny-Byfield et al 2014), whereas some gene pairs of maize have shown some largely uncorrelated patterns of expression.

In contrast, some studies have reported metabolic networks with absolute gene dosage when retaining homeologous gene pairs of *A. thaliana* in the long term when an absolute increase in the concentration of gene components is beneficial (Bekaert et al 2011). Interestingly, some researchers (Blanc & Wolfe 2004a, De Smet et al 2017, Wu & Qi 2010) identified duplicated genes in a network forming two parallel networks after polyploidy, each network including a homeolog from each duplicated gene pair. The expression of homeologs in different pathways was not seemingly correlated while each homeolog's expression was interrelated to the other components in the same pathway. This was referred to as concerted divergence of gene expression (Blanc & Wolfe 2004a).

4.2.2.2 Subgenome dominance

Another potential mode of expression divergence in allopolyploids discussed in previous studies is subgenome dominance, where one of the parental genomes (subgenomes) is preferentially expressed and along with the time, possesses greater gene retention compared to the other subgenome/s (Schnable et al 2011). Generally, severe chromatin modifications and low methylated transposable element (TE) abundance near genes have been associated with the dominant subgenome (Alger & Edger 2020, Liang & Schnable 2018). The other submissive subgenome undergoes fractionation to a greater extent (Bird et al 2018, Thomas et al 2006). This has been observed as true in the aspect of complexes and genetic pathways (Bird et al 2018, Cheng et al 2016, Edger et al 2019, Wang et al 2022b) such as, a certain subset of pathways is controlled by one subgenome, while the rest of the pathway is controlled by the other subgenome(s). For example, some metabolic pathways related to colour, flavor, and aroma in octoploid cultivated strawberries (Edger et al 2019) and glucosinolate biosynthesis pathway in allotetraploid *Brassica juncea* (Yang et al 2016) are largely dependent on the control of the dominant subgenomes identified. This would result in the partitioning of phenotypic traits to different subgenomes, as has been observed during fruit development in

blueberry (Colle et al 2019), production of long lint fibers in cotton (Samuel Yang et al 2006) and controlling grain size and form in wheat (Pfeifer et al 2014). While biased expression is the case for many polyploids, there are several examples of polyploids showing little or no bias but random gene loss or expression between homeologs (Burns et al 2021, Garsmeur et al 2014, Li et al 2019).

Thus, these are some theories that have been introduced to explain the expression divergence of duplicated genomes or genetic pathways after allopolyploidy. We are particularly interested in the trichome initiation pathway in this study, and therefore it is worth exploring the gene expression literature in this pathway. Unsurprisingly, the majority of the studies have been conducted in *Arabidopsis thaliana* as explained below. This species is also of particular interest because our study group is in the *Brassicaceae* system, *Pachycladon*, and *Arabidopsis thaliana* is evolutionarily relatively close to one of the subgenomes of our study system.

4.2.3 Gene expression studies of *Arabidopsis* trichome initiation pathway

The trichome initiation pathway described from *Arabidopsis thaliana* has been studied extensively in the past as discussed in the previous chapters. Hence, it is an ideal model for an exploration of the effects of whole genome duplication on the gene expression of a genetic pathway. Indeed, the most important role of the pathway is played by the central complex of *TTG1-GL1-GL3*/*EGL3*. The following is a short review of the knowledge gained in numerous previous studies done on the expression of these important components of the trichome initiation complex. To this end, the key transcription factors, associated hormonal mechanisms, and epigenetic modifications underlying trichome development in *Arabidopsis* have been addressed (Wang et al 2019). But here, we are especially interested in the expression regulation of key transcription factors.

TTG1 is found to be expressed ubiquitously in all the epidermal cells in *Arabidopsis* such as in leaves, stem, roots, and seed coats at all developmental stages and preferentially localized in cell nuclei (Li et al 2020, Zhao et al 2008). Hence, the transcription of *TTG1* is detected in all major organs of *Arabidopsis* (Walker et al 1999) reporting no enhanced

expression in trichomes (Bouyer et al 2008). These observations prove the fact that *TTG1* is pleiotropic and involved in multiple genetic pathways. The TTG1 protein is important for the appropriate bHLH distribution in the nuclei of young epidermal cells but is not essential for GL1 distribution though (ZHAO et al. 2008). *GLABRA2* (*GL2*), *TRANSPARENT TESTA GLABRA2* (*TTG2*), *CPC*, and *ETC1* are direct targets of TTG1. *GL2*, a homeodomain gene (HD-Zip,) and *TTG2*, a WRKY transcription factor, are also required for normal trichome development (Rerie et al 1994, Zhang et al 2003).

GL1 is also expressed ubiquitously in young leaf primordia and particularly very highly expressed in developing trichome cells. However, the expression of *GL1* ceases when trichomes start to initiate branches and become mature (Kirik et al 2005, Larkin et al 1993). *GL1*'s close paralog, *MYB23*, lacks the initial ubiquitous expression pattern and is expressed exclusively in developing trichomes (Kirik et al 2005). GL1 supports the interaction of GL3 in the TTG1-bHLH-GL1 complex with the promoters of specific downstream targets, and TTG1 helps the bHLH-GL1 interaction (Pattanaik et al 2014, Wei et al 2019, Zhao et al 2008). *TTG2* is also directly regulated by GL1 (Ishida et al 2007). However, over-expression of GL1 decreases the trichome identity (Wang et al 2019). GL1 and TTG1 do not interact directly, but they both interact with GL3/EGL3 separately to build the TTG1-GL1-GL3/EGL3 complex (Payne et al 2000, Zhao et al 2008).

GL3 and *EGL3* are most highly expressed in undifferentiated leaf primordia at a very young stage (Morohashi & Grotewold 2009, Patra et al 2013, Zhao et al 2008). During early leaf development, *GL3* is highly expressed, particularly in undifferentiated cells close to the basal edge of the leaf, while *EGL3* expression is more widespread, but high in trichome initials than surrounding epidermal cells (Patra et al 2013, Zhao et al 2008). When the leaves are more mature, strong *GL3* expression is only limited to trichomes while *EGL3* is expressed in trichomes and pavement cells (Balkunde et al 2010, Zhao et al 2008). Overall, *EGL3* exhibits a more widespread transcription pattern in the epidermal pavement cells than *GL3*, while *GL3* shows high levels of transcription particularly in trichomes (Patra et al 2013, Pattanaik et al

2014). Therefore, in *Arabidopsis*, these two genes, *GL3* and *EGL3*, have overlapping, yet distinct patterns of transcription (Morohashi et al 2007, Patra et al 2013). However, both *GL3* and *EGL3* are partially redundant in trichome initiation, as demonstrated by the double *gl3/egl3* mutant being completely glabrous, while the *gl3* mutant had much-reduced trichome initiation and branching than *egl3* (Zhang et al 2003). The downstream gene, *GL2* is a direct target of GL3 and EGL3 (Morohashi et al 2007). The transcriptional repression of *GL3* regulates a negative autoregulatory loop targeting its promoter, which is independent of TTG1 and GL1 (Morohashi et al 2007). This negative autoregulation might be happening in the trichome initiation cells after trichome initials have been initiated or in the cells destined not to become trichomes (Morohashi et al 2007, Zhang et al 2003). Further, the negative regulators *CPC* and *ETC1* are directly activated by GL3 and GL1 (Morohashi & Grotewold 2009, Morohashi et al 2007).

Negative regulation of the TTG1-GL1-bHLH complex is controlled by the R3 MYB negative regulators TRIPTYCHON (TRY) (Schellmann et al 2002), CAPRICE (CPC) (Wada et al 1997), and ENHANCER OF TRY and CPC1 and 2 (ETC1 and ETC2) (Kirik et al 2004a, Kirik et al 2004b). The R3 MYBs compete with the R2R3 MYB GL1 to bind with the bHLHs to build a negative regulatory complex that disrupts the function of the activating complex contributing to result in a glabrous phenotype in *gl1*-mutant background (Pattanaik et al 2014, Zhao et al 2008, Zheng et al 2016b, Zhou et al 2014).

4.2.4 Studies on duplicated trichome initiation pathway outside *Arabidopsis*

Many plant species produce epidermal outgrowths. In allotetraploid cotton (*Gossypium spp.*), the fiber cells on the seed coat are similar to trichomes and are controlled by a similar set of regulators as in *A. thaliana* (Wang et al 2019). Since Cotton is a polyploid, its trichome initiation pathway has been duplicated and has duplicated copies of the component genes. In *Gossypium arboretum* and *Gossypium hirsutum*, the fiber cell determination is controlled by the R2R3 MYB genes, *GaMYB2* and *GhMYB109*, which are highly expressed in the early stages of fiber cell development and interact with R-like bHLH proteins (GhMYC1) and WD-repeat

proteins (TTG1) suggesting multimeric complexes between MYB, bHLH, and WD repeat proteins trigger trichome formation in *Arabidopsis* and cotton (Suo et al 2003, Wang et al 2004b). But, in cotton, this complex has not been identified as a central complex as in *Arabidopsis* in the means of controlling the key *GL2* downstream gene for triggering trichome initiation, although the complex is being found responsible for fiber initiation or elongation (Wang et al 2019).

Similar R2R3 MYB transcription factors have been found in Asterids too. In *Antirrhinum majus* or Snapdragon (Plantaginaceae), the formation of a conical shape in petal epidermal cells, which are referred to as trichomes, is regulated by the R2R3 MYB-related transcription factor MIXTA (Glover et al 1998, Martin et al 2002, Noda et al 1994). Overexpression of *MIXTA* has induced the formation of multicellular trichomes and conical cells in *Nicotiana tabacum* (Solanaceae) and ectopic trichomes on anthers in *Solanum dulcamara* (Solanaceae) (Glover et al 2004, Glover et al 1998, Payne et al 1999). However, it is reported that neither R nor R-like bHLH nor WD repeat proteins play a role in multicellular trichome (and conical cell) formation in *Antirrhinum*, *Solanum*, or maize (Lloyd et al 1992, Payne et al 1999, Serna & Martin 2006).

4.2.5 Expression divergence of *Pachycladon* duplicated trichome initiation pathway

The expression data of *A. thaliana* key genes involved in the trichome initiation pathway from previous studies provide a head start to the study of the expression of the *Pachycladon* duplicated trichome initiation pathway, as both species belong to the Brassicaceae family. In addition, *Pachycladon*'s one subgenome is closely related to *A. thaliana* (Joly et al 2009, Zhao et al 2010). Moreover, *Pachycladon* can be regarded as an emerging model system, especially for gene duplication studies, because as stated in Section 1.3, it is derived from a hybridization event that happened between highly diverged progenitors followed by polyploidy and diversification. The species possess a wide range of morphological variation including trichome phenotype, which makes it more interesting to include different *Pachycladon* species with different trichome patterns in our study. The findings of the molecular evolution of the

Pachycladon trichome initiation pathway discussed in the previous chapter could be coupled with its expression divergence study in this chapter to build up a comprehensive account of the fates of gene duplicates in the pathway.

The *Pachycladon* polyploid genome is made of two very distantly related parental subgenomes (Joly et al 2009, Yogeeswaran et al 2011, Zhao et al 2010) and hence its trichome initiation pathway is composed of diverged duplicates. Therefore, this work provides the opportunity to explore the relevance of the models to the system such as gene dosage balance, absolute dosage, concerted divergence, or subgenome dominance.

We have generated the expression profiles of the four genes of interest (*TTG1*, *GL1*, *GL3*, and *EGL3*) involved in four *Pachycladon* species' trichome initiation complex, and for *P. cheesemanii*, we have produced expression results for two developmental stages/tissues (whole young seedlings and mature leaves). The questions we try to resolve in this chapter are, 1) What are the expression pattern that exists among the gene homeologs and species of *Pachycladon*? 2) Are there differences in expression for different developmental stages/tissues in *Pachycladon cheesemanii*? A gene expression study related to the trichome initiation complex in *Pachycladon* has not been done before and we are curious to see if that is similar to *A. thaliana* because both are in the family Brassicaceae. 3) How do sequence variation and molecular evolution (polymorphisms, disruptions, etc.) of the genes of interest relate to their expression patterns? With the findings of molecular evolution and gene expression of the trichome initiation complex of four *Pachycladon* species, we attempt to add insights to an underexplored area on duplicated gene fate in the context of a genetic pathway. Based on the sequence analysis results obtained in Chapter 3, Type 1 sequences were generally found intact and Type 2 sequences were likely fractionated, we hypothesize that the Type 1 subgenome has a biased expression.

4.3 MATERIALS AND METHODS

4.3.1 Plant materials collection, RNA extraction, and cDNA synthesis

Seeds of four *Pachycladon* species were collected from the field (Table 4.1) and were sown in autoclaved Daltons Premium Seed Mix soil, in 6×6 cm pots arranged in a fully randomized design (8 pots per species and about 9 seeds sown in each pot), watered, covered with aluminum foil and placed at 4 °C for 10 days to overcome seed dormancy. After that, the pots were placed in a growth room at 20 °C on a 16-hour light and 8-hour dark cycle. The initial plan was to compare expression patterns between homeologs of each gene within each species at two developmental stages; however, due to differences in germination rates, growth rates, and seed availability, not all species could be sampled at both time points. At 30 days post-germination, 2-3 whole seedlings without roots were collected per biological replicate from *P. enysii* and *P. cheesemanii* to analyze gene expression at all stages of trichome development, including the earliest stages of young leaf primordia. At approximately four months post-germination, 2-3 fully expanded young leaves were collected from *P. fastigiatum*, *P. novae-zealandiae*, and *P.cheesemanii* to analyze the gene expression when the trichomes are at the mature stage. (Figure 4.1 shows the plants collected from young *P.cheesemanii* and leaves collected from mature *P.cheesemanii*). The collected leaves were immediately flash-frozen in liquid nitrogen at the point of collection and then stored at -80 °C. Seedlings or leaves from three biological replicates were collected from each species. The reason for using seedlings for some species and fully expanded young leaves for others was the availability of the plants when ideally the idea was to collect plants/leaves from two growth stages as young and mature for all the species. As the outermost layer of the apical meristem differentiates into the epidermis and trichomes in young seedlings (Larkin et al 1996), and separating only the meristem was not feasible and not sufficient for RNA extractions, it was required to collect the whole seedling to assess the gene expression at the young growth stage. Comparatively, collecting a few fully expanded leaves to extract RNA from mature trichome cells was feasible

Table 4.1: *Pachycladon* species used in the study

Species	Population ID	Field location where the seeds were collected	Biological replicate	RNA Qubit concentration (ng/µL)	Labchip quality score	cDNA samples
Young *P.enysii*	Pch 104	Wye Creek, South Queenstown	P.en 1	21	7.2	Y-P.en-cDNA-1
			P.en 2	153	6.9	Y-P.en-cDNA-2
			P.en 3	30.8	7.6	Y-P.en-cDNA-3
Mature *P.fastigiatum*	Pch 879-5	Wye Creek, South Queenstown	P.fs 1	43	7.5	M-P.fs-cDNA-1
			P.fs 2	33.2	7.8	M-P.fs-cDNA-2
			P.fs 3	147	7.9	M-P.fs-cDNA-3
Mature *P.novae zealandiae*	Pch 873-X	Wye Creek, South Queenstown	P.nz 1	96.8	8.3	M-P.nz-cDNA-1
			P.nz 2	258	8.1	M-P.nz-cDNA-2
			P.nz 3	173	8.2	M-P.nz-cDNA-3
Young *P.cheesemanii*	Pch 875-5	Wye Creek, South Queenstown	Y-P.ch 1	306	7.3	Y-P.ch-cDNA-1
			Y-P.ch 2	266	7.3	Y-P.ch-cDNA-2
			Y-P.ch 3	236	7.1	Y-P.ch-cDNA-3
Mature *P.cheesemanii*	Pch 875-5	Wye Creek, South Queenstown	M-P.ch 1	298	7.2	M-P.ch cDNA-1
			M-P.ch 2	128	8.2	M-P.ch cDNA-2
			M-P.ch 3	175	7.4	M-P.ch cDNA-3

and sufficient for RNA extraction. We could not collect leaves from *P. enysii* in the mature stage because not enough plants were left for it after the whole seedlings were collected in the young stage. *P. fastigiatum* and *P. novae-zealandiae* had very low germination rates and rather slow growth compared to the other species, so the seedlings were not collected in the young stage because of inadequate plant materials and leaves were collected for assessing gene expression at the mature stage. But, comparatively, *P. cheesemanii* seedlings had higher germination rates and fast growth enabling collection of plants/leaves at both young and mature growth stages.

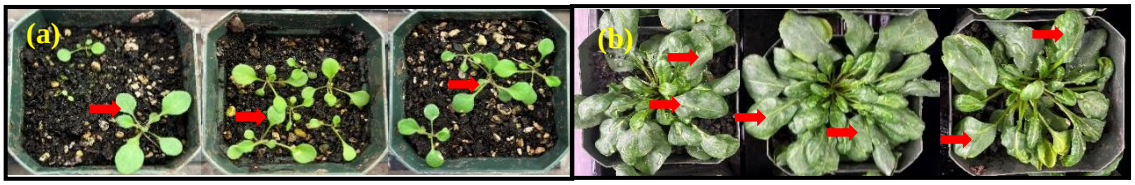


Figure 4.1: *Pachycladon cheesemanii* plants and leaves collected for RNA extraction (a) Pch 875-5 one-month-old young whole seedlings and (b) Pch 875-5 four months old mature fully expanded leaves, shown by red arrows were collected from three biological replicates for each experiment condition.

RNA extractions from frozen material were completed using a Qiagen RNeasy Plant Mini Kit (cat. no. 74904) following the manufacturer's instructions, including the digestion of any potentially contaminating genomic DNA using the DNase treatment. RNA concentrations were measured using the Qubit RNA BR (Broad Range) assay kit (cat. no. Q10210) (Qubit readings are listed in Table 4.1). Further, RNA integrity and quality were analysed by the RNA standard LabChip assay (PerkinElmer LabChip GX Touch HT instrument at Massey Genome Services).

Next, for each sample, first strand synthesis of the complementary DNA (cDNA) was done from RNA using Invitrogen SuperScript III First-Strand Synthesis System for RT-PCR (cat. no. 18080-051) following the kit instructions using the provided 50 µM oligo$(dT)_{20}$ primers, because consistent production of cDNA from full-length mRNA was preferred. Because of variability in the RNA available among samples, the RNA amount used to synthesize cDNA was consistent among biological replicates within a species but varied among species. As the focus here is on the differential expression between homeologs of a gene within a species, the impact of this variation is minimal; however, it meant that meaningful comparisons of expression levels among species were not possible. To control for potentially contaminating genomic DNA (even though a DNAse digestion was performed during RNA extractions), for all biological replicates, a reverse transcriptase amplification control sample (no RT control) was included; these samples were treated the same as all other samples except no reverse transcriptase enzyme was added. The cDNA for all samples was diluted to a final concentration of 2.5 ng/µL for use in qPCR reactions. cDNA concentrations were calculated according to the RNA concentration and volume used to make the cDNA.

4.3.2 Selection of qPCR reference genes

Reference genes assessed for expression stability in different plant tissues and organs at different developmental stages were selected from the literature. Because *Arabidopsis thaliana* is the best-studied close relative of *Pachycladon*, *A. thaliana* studies were mainly used to identify reference genes for this study (Czechowski et al 2005, Dekkers et al 2011, Jin et al

2019); however information from other species was also considered (Expósito-Rodríguez et al 2008). From these studies, three stably expressed reference genes were selected (Table 4.2).

Table 4.2: Selected reference genes for the study

Gene symbol	**Gene name**	***Arabidopsis* gene**
PPR	Pentatricopeptide repeat (PPR) superfamily protein	At5G55840
CAC	Clathrin adaptor complexes medium subunit family protein	AT5G46630
TIP41	TOR signaling pathway family protein	AT4G34270

4.3.3 qPCR SYBR Green primer design

Homeolog-specific qPCR primers were designed considering specific polymorphisms in the sequence assemblies of four genes of interest: *TTG1*, *GL1*, *GL3* and *EGL3*. Primers were designed using Primer3web version 4.1.0 online software (Untergasser et al 2012) and also manually in Geneious (Geneious 9.1.8 2017).

For each gene, all available sequences across homeologs and species were aligned using the Geneious multiple alignment method. The FASTA formatted alignments were uploaded to the Primer3web to identify left (forward) and right (reverse) primers based on several criteria. Critically, conserved polymorphisms that differentiate homeologs for each gene across all species were identified from the gene alignments and were used during primer design. The optimal parameters for primer design were selected from the literature (Bustin & Huggett 2017, Thornton & Basu 2011). Nevertheless, following the procedure, Primer3web often generated primers that were not homeolog-specific or contained single polymorphism in the middle of the primers rather than at the 3’ end or close to 3’ end which are necessary for differentiating homeologs. Hence, manual primer design was performed in Geneious 9.1.8 considering the following criteria: at least one homeolog-differentiating polymorphism at the 3’ end (or within 2-3 bases from the 3’ end), primer length 18-24 bp, primer melting temperature (T_m) ~63-64 °C with primer pair T_m difference < 2 °C, product size between 80-200 bp, mono- and di-nucleotide repeats maximum 4, 3’ stability with 1-2 GC bases and GC percentage 35-

70%. Primers were designed in exon-exon junctions whenever possible to avoid amplification of any contaminating genomic DNA in cDNA samples. Finally, the designed primers were checked for secondary structures and primer dimers using the Integrated DNA Technologies (IDT) oligoanalyzer tool (https://sg.idtdna.com/calc/analyzer (Owczarzy et al 2008)). If IDT oligoanalyzer revealed primer hairpin T_m close to the primer Tm (~60 °C) and if the Gibbs free energy (ΔG) of primer secondary structure is less than –9 for self-dimers and hetero-dimers, then those primers were rejected and new primers were designed. Next, amplicon secondary structures were analysed using IDT UNAFold tool (https://sg.idtdna.com/Unafold/ (Owczarzy et al 2008)). Amplicons were accepted only if their T_m was much lower than the qPCR annealing temperature (60 °C). Lastly, for primer validation, primer specificity was checked in NCBI Primer-Blast selecting the database nr (non-redundant) to ensure the primer is specific for the gene of interest and other nonspecific genes are not amplified (Ye et al 2012). The selected primers are listed in Table 4.3 below.

Table 4.3: qPCR primers used. 'Ref' indicates the reference genes primers.

Gene	Homeolog	Primer	Sequence 5'-3'	T_m (°C)	Amplicon size (bp)
TTG1	Type 1	1920-Pch-TTG1T1-F 1922-Pch-TTG1T1-R	GAAGATTCTTCAACCGTTGAGCCTA CCCAAATCGTACAAGTCGTATCAATAC	63.61 62.8	148
	Type 2	1923-Pch-TTG1T2-F 1925-Pch-TTG1T2-R	CTTCCCTGCATCGTCCCTCG CGCTCGTCTTACTGTTGTTGAGG	65.09 64.13	129
GL1	Type 1	1994-Pch-GL1T1-F 2004-Pch-GL1T1-R	CGCCACCACAACTTCTTCTCT TGATTCCATAAAGCGTTATTGCCATAAT	63.02 62.84	192
	Type 2	1996-Pch-GL1T2-F 2006-Pch-GL1T2-R	ACAAGACACGAATATCTGCAACAACAC GAACTCGTCTTCATGCACCCAG	65.08 63.96	175
GL3	Type 1	2235-Pch-GL3T1-F 2236-Pch-GL3T1-R 2237-Pch.nz-GL3T1-R	AGCTCATGGACGGCGAGC CGTCTTTACATTTCTCTGCTGCTCT ACGTCTTTACATTTCTTTGCTGCTCT	62.5 61.4 61.7	170 171
	Type 2	2292-Pch-GL3T2-F 2293-Pch-GL3T2-R	GGAGTACTAGAATGGCGAGATGGATAC GCTCGTCTGGTTACCTAAGATCCG	62.2 62.6	170
EGL3	Type 1	2252-Pch-EGL3T1-F 2253-Pch-EGL3T1-R	GCAGTTTGTGGGCAAAGAACTT TCATCGTCCATATTTACATGGTTACTC	60.4 59.1	177
	Type 2	2254-Pch-EGL3T2-F 2255-Pch-EGL3T2-R	CCGATAGTAAAGTCTTCACTCGTTCT CGTGCCGATCTCAAGGACG	60.4 60.9	101
Ref-PPR	Type 1 and Type 2	2003-Pch-Ref-PPR-F 2012-Pch-Ref-PPR-R	GAACCACAGATTCCAGGAGTCT CCGTTAAGCAGACCAATGTACTTTC	61.96 62.87	96
Ref-CAC	Type 1 and Type 2	1928-Pch-Ref-CAC-F 1930-Pch-Ref-CAC-R	CCAGGACAAACAGAGTCAACACT GAACTCGCAAACCAGATGCTG	63.2 62.55	135
Ref-TIP41	Type 1 and Type 2	1929-Pch-Ref-TIP41-F 1931-Pch-Ref-TIP41-R	TTCAGGAGCAAGCCGTCCCA GTGAGAGAGCAGCCAAATCAATCTG	66.48 64.66	184

Next, the primers were tested on all *Pachycladon* cDNA and genomic DNA templates using FastStart Taq DNA polymerase (Roche) using 10X PCR buffer with $MgCl_2$ 1.0 µL, 10 mM dNTPs 0.2 µL, 10 µM forward and reverse primers 0.5 µL each, DMSO 0.3 µM, FastStart Taq 0.08 µL, cDNA template 1.0 µL and water up to 10 µL reaction volume. The reactions were incubated at 95 °C for 4 mins, then 30 cycles of 95 °C for 30 s, 60 °C for 30 s 72 °C for 12 s, and then 72 °C for 7 mins. After that, the PCR products were sequenced using Sanger sequencing to confirm that they are the expected homeolog specific PCR products before qPCR experiments were performed with the primers.

4.3.4 Assay validation

Before the actual experiments were set up and valuable experimental samples consumed, it was necessary to carefully evaluate the qPCR assay performance for specificity and efficiency.

4.3.4.1 Specificity analysis

As stated in 4.3.3, primer specificity was initially checked using both genomic and cDNA by confirming PCR expected product size on agarose gels. Those products were then evaluated by Sanger sequencing. Further, the product specificity was precisely confirmed by qPCR melt curve analysis using cDNA as stated below in 4.3.4.2, to confirm that there were no nonspecific amplifications.

4.3.4.2 Efficiency analysis

Efficiency analyses were completed for each qPCR primer pair by generating standard curves against a template dilution series. Theoretically, in successive PCR cycles, the product should be doubled in amount, assuming the reaction has 100% amplification efficiency. But, in reality, this is not the case, and amplification efficiency is reduced due to target length, target sequence, primer sequence, buffer conditions, impurities present in the sample, cycling conditions, enzyme used and even pipetting accuracy (Arezi et al 2003).

Validated *P. enysii* PCR products from 4.3.3 for each primer pair were taken as templates for making dilution series. From the PCR product for each reference gene, 10^4, 10^6,

$10^{8,}$ and 10^{10} -fold dilutions were made and, for genes of interest 10^2, 10^4, 10^6, 10^8 fold dilutions were made to test primer pair efficiencies. Reference genes are more abundant in the genome and are generally expressed higher than the genes of interest, so higher fold dilutions were selected for the reference genes. Otherwise, they may result in very low Cq values, that would result in outliers from the standard curves generated in the qPCR run. Two replicates for each dilution and two negative controls for each primer pair were included in the validation design. Then 10 µL qPCR reactions were prepared using Roche LightCycler 480 SYBR Green I Master mix (cat. no. 04707516001) following the kit instructions using 5 µM forward and reverse primer pairs (1.0 µL from each 5 µM forward and reverse primers, 5.0 µL from the master mix, 2.5 µL from the template and 0.5 µL water to top up to 10 µL total reaction volume). The reactions were run using the qPCR program shown in Table 4.4 in the Roche LightCycler 480 machine according to the instrument manual. The standard curves were obtained for each primer pair through the LightCycler 480 software release 1.5.1.62 SP3, and the reaction efficiencies were also calculated in the software by analysing the slope of the log-linear portion of the standard curve. PCR efficiency = $10^{-1/slope}$, when the logarithm of the initial template concentration (the independent variable) is plotted on the x axis and Cp (Crossing point, the dependent variable) is plotted on the y axis (Kontanis & Reed 2006, Roche 2001). The melt curve readings were also run in the PCR program (Table 4.4) to check for primer specificity.

After efficiency analysis of primer pairs, another dilution series was made using the unamplified original cDNA templates to make sure they don't carry inhibitors that slow down the reaction efficiency, at the actual low dilutions that will be used in the experiments. Potential inhibitory substances are endogenously present in biological tissues and may accumulate during RNA extraction and reverse transcription reactions. Therefore, a cDNA dilution series was performed for all samples against one of the primer pairs with good reaction efficiency (TTG1-T1-1920F×1922R). Depending on the starting RNA concentrations and available RNA volumes from species, the cDNA dilution factors differed by species, but the cDNA samples were

diluted to have nearly similar final concentrations (2.5, 3.3, 5 and 10 ng/µL) in each sample to test for their efficiencies (Table A4.1).

Table 4.4: qPCR program used in the LightCycler 480

Setup					
Block Type			Reaction volume (µL)		
96			10		
Detection format		Excitation filter		Emission filter	
SYBR Green		465		510	
Programs					
Program Name		Cycles		Analysis mode	
Pre-Incubation		1		None	
Amplification		45		Quantification	
Melting Curve		1		Melting curve	
Cooling		1		None	
Temperature targets					
	Target (°C)	Acquisition Mode	Hold (hh:mm:ss)	Ramp rate (°C/s)	Acquisitions (per °C)
Pre-Incubation	95	None	00:10:00	4.4	-
Amplification	95	None	00:00:10	4.4	-
	60	None	00:00:20	2.2	-
	72	Single	00:00:10	4.4	-
Melting Curve	95	None	00:00:05	4.4	-
	65	None	00:01:00	2.2	-
	97	Continuous	-	-	5
Cooling	40	None	00:00:10	1.5	-

4.3.5 qPCR experiments

Following the validation experiments, real-time quantification experiments were carried out in 96-well plates for five templates (young *P. enysii*, young *P. cheesemanii*, mature *P. cheesemanii*, mature *P. fastigiatum*, and mature *P. novae-zealandiae*; a complete set as young and mature for each species was unable to process due to the availability of plants as described in 4.3.1) using the primer combinations shown in Table 4.3 to measure the homeolog expression of the four genes of interest. 2.5 ng/µL concentrated cDNA working stocks were used with the reaction conditions discussed below.

Each reaction set per primer pair included three technical replicates per biological replicate, three biological replicates per species, duplicate reactions for -RT templates, two negative water controls, two technical replicates per reference gene: likewise three reference genes per biological replicate, and duplicate reactions for reference gene negative controls. One such experimental setup for *P. enysii TTG1* Type 1 and Type 2 homoeologs is shown in Figure 4.2 as an example.

96 well plate		Y-P.en-cDNA-1 template			Y-P.en-cDNA-2 template			Y-P.en-cDNA-3 template				Negative water	
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
TTG1 -T1 GOI	**A**	Y-P.en cDNA-1	Y-P.en-cDNA-1	Y-P.en-cDNA-1	Y-P.en-cDNA-2	Y-P.en-cDNA-2	Y-P.en-cDNA-2	Y-P.en-cDNA-3	Y-P.en-cDNA-3	Y-P.en-cDNA-3		NTC	NTC
TTG1 -T2 GOI	**B**	Y-P.en-cDNA-1	Y-P.en-cDNA-1	Y-P.en-cDNA-1	Y-P.en-cDNA-2	Y-P.en-cDNA-2	Y-P.en-cDNA-2	Y-P.en-cDNA-3	Y-P.en-cDNA-3	Y-P.en-cDNA-3		NTC	NTC
No RT control (*TTG1* -T1)	**C**	Y-P.en-cDNA-1 no RT	Y-P.en-cDNA-1 no RT		Y-P.en-cDNA-2 no RT	Y-P.en-cDNA-2 no RT		Y-P.en-cDNA-3 no RT	Y-P.en-cDNA-3 no RT				
Ref gene 1 *CAC*	**D**	Y-P.en-cDNA-1	Y-P.en-cDNA-1		Y-P.en-cDNA-2	Y-P.en-cDNA-2		Y-P.en-cDNA-3	Y-P.en-cDNA-3			NTC	NTC
Ref gene 2 *TIP41*	**E**	Y-P.en-cDNA-1	Y-P.en-cDNA-1		Y-P.en-cDNA-2	Y-P.en-cDNA-2		Y-P.en-cDNA-3	Y-P.en-cDNA-3			NTC	NTC
Ref gene 3 *PPR*	**F**	Y-P.en-cDNA-1	Y-P.en-cDNA-1		Y-P.en-cDNA-2	Y-P.en-cDNA-2		Y-P.en-cDNA-3	Y-P.en-cDNA-3			NTC	NTC
	G												
	H												

Figure 4.2: Example experimental setup (96 well plate) for *P. enysii* for *TTG1* Type 1 and Type 2 homeologs. Y-P. en-cDNA-1,2 and 3 are the biological replicates of the species. *TTG1* Type 1 and *TTG1* Type 2 are the homeologs (genes) of interest (GOI). Three technical replicates were included for each biological replicate. NTC stands for No Template Control, which is the negative water control and at least duplicate reactions were included for each primer pair used. No reverse transcriptase (No RT) negative controls were checked against one of the selected primer pairs only, because if no RT controls are amplified with one primer pair, then it is assumed there are not genomic DNA contaminations in the starting RNA sample, which means no amplification is expected with other primers as well. Three reference genes were used to normalize the results: *CAC*, *TIP41* and *PPR*. One of the samples with one reference gene (P.en cDNA 2 with *TIP41*) was used as the inter-plate control sample to run in all the plates, to normalize potential inter-plate variation.

The same qPCR recipe found in the LightCycler 480 SYBR Green I Master kit used for primer efficiency tests above, was used in the experiments. Also, the same qPCR program (Table 4.4) used for primer efficiency tests was used in the final experiments. The resulting Cp values from each experiment were exported for analysis.

4.3.6 Data analysis and relative quantification

4.3.6.1 Data pre-processing

The GenEx 7.0 standard version software (MultiD_Analyses_AB 2019) was used for qPCR data analysis. First, an input table of sample Cp values was prepared according to the software requirements and processed for the analysis (Table A4.2 in the Supplementary data). Data analysis was composed of two stages: data pre-processing and statistical analysis. All steps were carried out based on the software guidelines.

In the data pre-processing, first, inter-plate calibration was done to normalize the Cp values measured in different runs by adding the classification columns 'Inter-plate Control (IPC)' and plate number using the interplate control sample P.en cDNA 2 with *TIP41* reference gene. Then, the data was efficiency-corrected based on the specific primer efficiency values obtained in the efficiency tests above. Next, before averaging the Cp values of technical replicates, outliers were excluded. If an outlier occurred in a set of three technical replicates, those sets were identified by a standard deviation value more than 0.35, which corresponds to a maximum allowed difference of 0.5 cycles between the technical replicates (Hellemans & Vandesompele 2011). This threshold is applicable for Cp values in the range of 9-30. But when higher Cp values were recorded such as from 31-42, they are naturally expected to have more "noise" or variability associated with them, so can be expected to have up to a 1.5 difference in Cp values between technical replicates. Thus, after identifying a set of qPCR repeats of technical replicates that include outliers, first their melt peaks were analysed to see if there were any abnormal peaks resulting from nonspecific amplifications, but none was found. Then, Grubb's test was also used to exclude potential outliers. The potential outliers were identified when the value associated with that data point was larger than an upper limit of 1.5 times the inter-quartile range from the 75% percentile of value of the median of the dataset. Similarly, lower level of detection was set to 1.5 times the inter-quartile range down from the 25% percentile of value. Using these methods, no outlier data points were identified.

The qPCR technical repeats were then averaged, and expression of target genes was normalized to the expression of the three reference genes by the software using the following equation, where GOI stands for Gene of Interest and RG for Reference Gene (MultiD_Analyses_AB 2019):

$$Cpnorm = CpGOI - \frac{1}{n}\sum_{i=1}^{n} CpRGi$$

Next, relative quantification was performed for each column of gene of interest against the sample showing the highest expression (lowest Cp value) within each gene, across all the biological duplicates from all species. Lastly, the data pre-processing was completed by converting all linear values to log 2 values. This is a requirement to proceed to statistical data analysis because statistical methods assume that data is normally distributed, and gene expression data often tends to be normally distributed when expressed in logarithmic scale rather than in linear scale.

4.3.6.2 Statistical analysis of data

A t-test was done to analyse if there is a significant difference in the mean between each two groups of data (e.g., *P. enysii TTG1* Type 1 vs Type 2, *P. fastigiatum GL1* Type 1 vs Type 2, etc). Unpaired, 2-tailed t-tests were run for each analysis between the Type 1 and Type 2 groups against all genes. The same test was done to analyse the expression difference of each type in *P. cheesemanii* young and mature growth stages. Since more than one gene was tested at the same time (multiple testing), a corrected threshold p-value of 0.01274 was used according to Dunn-Bonferroni correction method, instead of default threshold p-value of 0.05.

4.3.6.3 Gene expression results validation using qRAT

The raw data was also analyzed using qRAT (qPCR - Relative Expression Analysis Tool) (Flatschacher et al 2022) for the validation of the results obtained using GenEx. The interplate- calibrated raw Cp values were input to the software as txt file. Then the data quality was assessed by the software using the parameters such as maximum replicate variability and a range of acceptable Cp values (between 5 and 35). After that, the technical replicates and

biological replicates were averaged, and processed for relative quantification using $\Delta Cq = Cq$ (target) – Cq (reference), where arithmetic mean of the Cq values of the reference genes is used when multiple reference genes involve. The statistical analysis was performed in the software including t-test, p value and standard error. The relative quantification was presented as a Fold Change (FC); $FC = 2^{-\Delta\Delta Cq}$, where Type 2 homeolog expression is calculated relative to the expression of Type 1 homeolog. The results are shown in Table A4.3 and the generated plots are shown in Figure A4.5.

4.4 RESULTS

4.4.1 qPCR assay primer efficiencies

For each gene of interest, homeolog specific gene expression was measured using a single homeolog-specific primer pair for each of the two homeologs. All assessments showed that primer efficiencies were between the range of 1.8-2.0 (Table 4.5), which is analogous to a primer efficiency from 90% to 100%. It is assumed the assays are robust and precise (Bustin et al 2009) when the efficiency of primers is 90% - 100%. The standard curves which were used to calculate the primer efficiencies and primer melting curves showing primer specificities are presented in the supplementary data (Figures A4.1 and A.4.2). It is important to have comparable efficiencies for statistical analysis to be accurate, when normalizing the sample Cp values with reference gene Cp values and also when comparing Cp values of gene homeologs, although certain analyses can account for small differences also (e.g., GenEx).

Table 4.5: qPCR primer efficiencies

Gene	Primer pair	Primer efficiency
TTG1-T1	1920F×1922R	1.974
TTG1-T2	1923F×1925R	1.924
GL1-T1	1994F×2004R	1.898
GL1-T2	1996F×2006R	1.93
GL3-T1	2235F×2236R	1.907
GL3-T1	2235F×2237R	1.884
GL3-T2	2292F×2293R	1.888
EGL3-T1	2252F×2253R	1.866
EGL3-T2	2254F×2255R	1.862
Ref1-*CAC*	1928F×1930R	1.937
Ref2-*TIP41*	1929F×1931R	1.908
Ref3-*PPR*	2003F×2012R	1.863

The cDNA efficiency analysis that was carried out for all five *Pachycladon* templates resulted in linear standard curves (Supplementary data Figure A4.3), indicating no significant inhibitory effect is present in the used dilution series. The dilution series results, concentrations, and primer efficiencies for the templates are also presented in the supplementary data (Table A4.1). Based on the assay validation, the lowest concentration used in the dilution series - 2.5 ng/µL - was found to be sufficient and was used as the working stock concentration for all qPCR experiments.

4.4.2 qPCR data statistical analysis

Table 4.6 shows the output (including t-tests) from the statistical analysis in GenEx that identifies differences in the expression levels between two homeologs of a given gene in each species/tissue tested; The graphs in Figure 4.3 show the difference of homeolog relative expression in log scale (Type 1 log expression - Type 2 log expression), where positive values indicate greater expression of Type 1 than Type 2 homeologs, and negative values indicate greater expression in Type 2 than Type 1 homeologs.

Table 4.6: Unpaired 2-tail t-test results between all GOI (*TTG1*, *GL1*, *GL3* and *EGL3*) Type 1 and Type 2 homeologs of (a) *P. enysii* (*P.en*) seedling, (b) *P. cheesemanii* (*P. ch*) seedling, (c) *P. cheesemanii* (*P. ch*) mature leaf, (d) *P. fastigiatum* (*P.fs*) mature leaf and (e) *P. novae-zealandiae* (*P.nz*) mature leaf. The first three rows of all tables show pre-processed log Cp measurements, then below them the number of measurements (Count), the average measurement (Mean), the standard deviation from the mean (STDEV), the degrees of freedom (df), the test statistic (t), the probability (p-value), standard error of the mean (SEM) and t-value (t*). The difference in A (Type 1) -B (Type 2) and confidence levels (CI) are shown in the graphs in Figure 4.3. The important 'P-value' row is colored in grey and further, the significant P-values less than 0.01274 (corrected P-value threshold for multiple testing) are highlighted in yellow, and also the P-values less than the default threshold p-value of 0.05 are highlighted in blue.

(a)

***P. enysii* seedling**	**GOI *TTG1* homeolog Type 1**	**GOI *TTG1* homeolog Type 2**	**GOI *GL1* homeolog Type 1**	**GOI *GL1* homeolog Type 2**	**GOI *GL3* homeolog Type 1**	**GOI *GL3* homeolog Type 2**	**GOI *EGL3* homeolog Type 1**	**GOI *EGL3* homeolog Type 2**
	0.46441	0.48457	4.11242	3.62222	6.33586	4.56798	1.10834	2.31718
	0.58999	0.50702	5.93826	4.53079	6.9142	4.49239	2.32939	3.99885
	0.43457	0.6459	5.55054	4.15796	7.07445	4.84573	1.93027	3.26407
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.49632	0.54583	5.20041	4.10366	6.77484	4.63537	1.78933	3.19337
STDEV	0.08248	0.08739	0.96196	0.45671	0.38851	0.18606	0.6226	0.84306
df		4.		4.		4.		4.
SD^2		0.00722		0.56698		0.09278		0.5492
t		0.7136		1.7839		8.60247		2.32039
P-Value		0.514901611		0.149005308		0.001003492		0.081102394
Confidence level		0.95		0.95		0.95		0.95
SEM		0.06938		0.6148		0.2487		0.60509
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		-0.04951		1.09675		2.13947		-1.40403
CI start/end (log)	-0.24212	0.14311	-0.61022	2.80372	1.44896	2.82998	-3.08403	0.27596
Fold change		-1.03491		2.13872		4.406		-2.64641
CI start/end (linear)	-1.18273	1.10428	-1.52649	6.98237	2.7301	7.11066	-8.47977	1.2108

(b)

***P. cheesemanii* seedling**	**GOI *TTG1* homeolog Type 1**	**GOI *TTG1* homeolog Type 2**	**GOI *GL1* homeolog Type 1**	**GOI *GL1* homeolog Type 2**	**GOI *GL3* homeolog Type 1**	**GOI *GL3* homeolog Type 2**	**GOI *EGL3* homeolog Type 1**	**GOI *EGL3* homeolog Type 2**
	0.4792	0.68471	7.73249	5.61657	8.04339	1.64603	4.88473	4.87682
	0.59855	0.53379	7.61713	4.98482	8.02145	1.93981	5.05133	5.1727
	0.45237	0.43406	7.31387	4.31619	7.83598	0.79412	5.05084	4.90021
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.51004	0.55085	7.5545	4.97253	7.96694	1.45998	4.99563	4.98324
STDEV	0.07782	0.12619	0.21622	0.65028	0.11394	0.59507	0.09605	0.16449
df		4.		4.		4.		4.
SD^2		0.01099		0.23481		0.18355		0.01814
t		0.47679		6.52594		18.60154		0.11267
P-Value		0.658388011		0.002847595		0.000049163		0.915719509
Confidence level		0.95		0.95		0.95		0.95
SEM		0.0856		0.39565		0.34981		0.10997
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		-0.04081		2.58197		6.50696		0.01239
CI start/end (log)	-0.27847	0.19684	1.48348	3.68046	5.53573	7.47818	-0.29295	0.31773
Fold change		-1.02869		5.98757		90.94709		1.00863
CI start/end (linear)	-1.21291	1.14619	2.79622	12.82125	46.38976	178.30168	-1.22514	1.24637

(c)

***P.cheesemanii* mature leaf**	**GOI *TTG1* homeolog Type 1**	**GOI *TTG1* homeolog Type 2**	**GOI *GL1* homeolog Type 1**	**GOI *GL1* homeolog Type 2**	**GOI *GL3* homeolog Type 1**	**GOI *GL3* homeolog Type 2**	**GOI *EGL3* homeolog Type 1**	**GOI *EGL3* homeolog Type 2**
	0.54979	0.90384	0.12055	1.78621	7.91427	0.	5.04682	3.32294
	0.38695	0.92658	1.86854	2.77365	8.75839	0.32372	4.99428	3.55537
	0.1373	0.76765	3.62651	3.36087	8.79224	2.10157	5.72605	4.65035
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.35802	0.86602	1.87187	2.64024	8.4883	0.80843	5.25571	3.84289
STDEV	0.20776	0.08595	1.75299	0.79576	0.49741	1.13153	0.40817	0.70887
df		4.		4.		4.		4.
SD^2		0.02528		1.8531		0.76389		0.33455
t		3.91345		0.69131		10.76181		2.9916
P-Value		0.017343278		0.527399847		0.000422683		0.040274337
Confidence level		0.95		0.95		0.95		0.95
SEM		0.12981		1.11148		0.71362		0.47226
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		-0.50801		-0.76838		7.67987		1.41283
CI start/end (log)	-0.86842	-0.1476	-3.85435	2.3176	5.69854	9.6612	0.10161	2.72404
Fold change		-1.42209		-1.70335		205.0556		2.66258
CI start/end (linear)	-1.82566	-1.10772	-14.46359	4.98501	51.93149	809.67829	1.07297	6.60722

(d)

***P. fastigiatum* mature leaf**	**GOI *TTG1* homeolog Type 1**	**GOI *TTG1* homeolog Type 2**	**GOI *GL1* homeolog Type 1**	**GOI *GL1* homeolog Type 2**	**GOI *GL3* homeolog Type 1**	**GOI *GL3* homeolog Type 2**	**GOI *EGL3* homeolog Type 1**	**GOI *EGL3* homeolog Type 2**
	0.71806	0.	2.36349	2.05252	7.66498	4.53701	0.53146	1.07389
	0.29386	0.13229	4.45696	0.	8.45027	3.95449	1.53774	2.43543
	0.50115	0.20411	3.45394	1.48735	8.35687	4.58089	2.58074	2.59139
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.50436	0.11213	3.4248	1.17996	8.15737	4.35746	1.54998	2.03357
STDEV	0.21212	0.10354	1.04704	1.06023	0.42897	0.34968	1.02469	0.83476
df		4.		4.		4.		4.
SD^2		0.02786		1.11018		0.15315		0.87341
t		2.87816		2.60936		11.89231		0.63374
P-Value		0.045096673		0.059460086		0.000286343		0.560676595
Confidence level		0.95		0.95		0.95		0.95
SEM		0.13628		0.8603		0.31953		0.76307
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		0.39223		2.24484		3.79991		-0.48359
CI start/end (log)	0.01386	0.77059	-0.14375	4.63343	2.91276	4.68706	-2.60221	1.63503
Fold change		1.31242		4.73984		13.92794		-1.39822
CI start/end (linear)	1.00965	1.70597	-1.10477	24.8199	7.53059	25.75996	-6.07215	3.10593

(e)

***P. novae-zealandiae* mature leaf**	**GOI *TTG1* homeolog Type 1**	**GOI *TTG1* homeolog Type 2**	**GOI *GL1* homeolog Type 1**	**GOI *GL1* homeolog Type 2**	**GOI *GL3* homeolog Type 1**	**GOI *GL3* homeolog Type 2**	**GOI *EGL3* homeolog Type 1**	**GOI *EGL3* homeolog Type 2**
	1.53506	1.54343	3.80791	1.70847	9.07585	5.2865	1.7998	1.72717
	1.59496	1.42853	3.2362	2.61885	8.94254	5.91741	2.29265	0.29905
	1.39152	1.29837	2.62475	1.50514	8.90969	5.52077	2.3393	0.
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	1.50718	1.42344	3.22295	1.94415	8.97603	5.5749	2.14392	0.67541
STDEV	0.10455	0.12261	0.59169	0.59308	0.088	0.31892	0.29892	0.92304
df		4.		4.		4.		4.
SD^2		0.01298		0.35092		0.05473		0.47068
t		0.90009		2.64388		17.80631		2.62155
P-Value		0.418962054		0.057347518		0.000058449		0.058703549
Confidence level		0.95		0.95		0.95		0.95
SEM		0.09303		0.48368		0.19101		0.56017
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		0.08373		1.2788		3.40113		1.46851
CI start/end (log)	-0.17455	0.34202	-0.06412	2.62172	2.87081	3.93145	-0.08676	3.02378
Fold change		1.05976		2.42637		10.56433		2.76736
CI start/end (linear)	-1.12862	1.26753	-1.04545	6.15481	7.31475	15.25754	-1.06199	8.13297

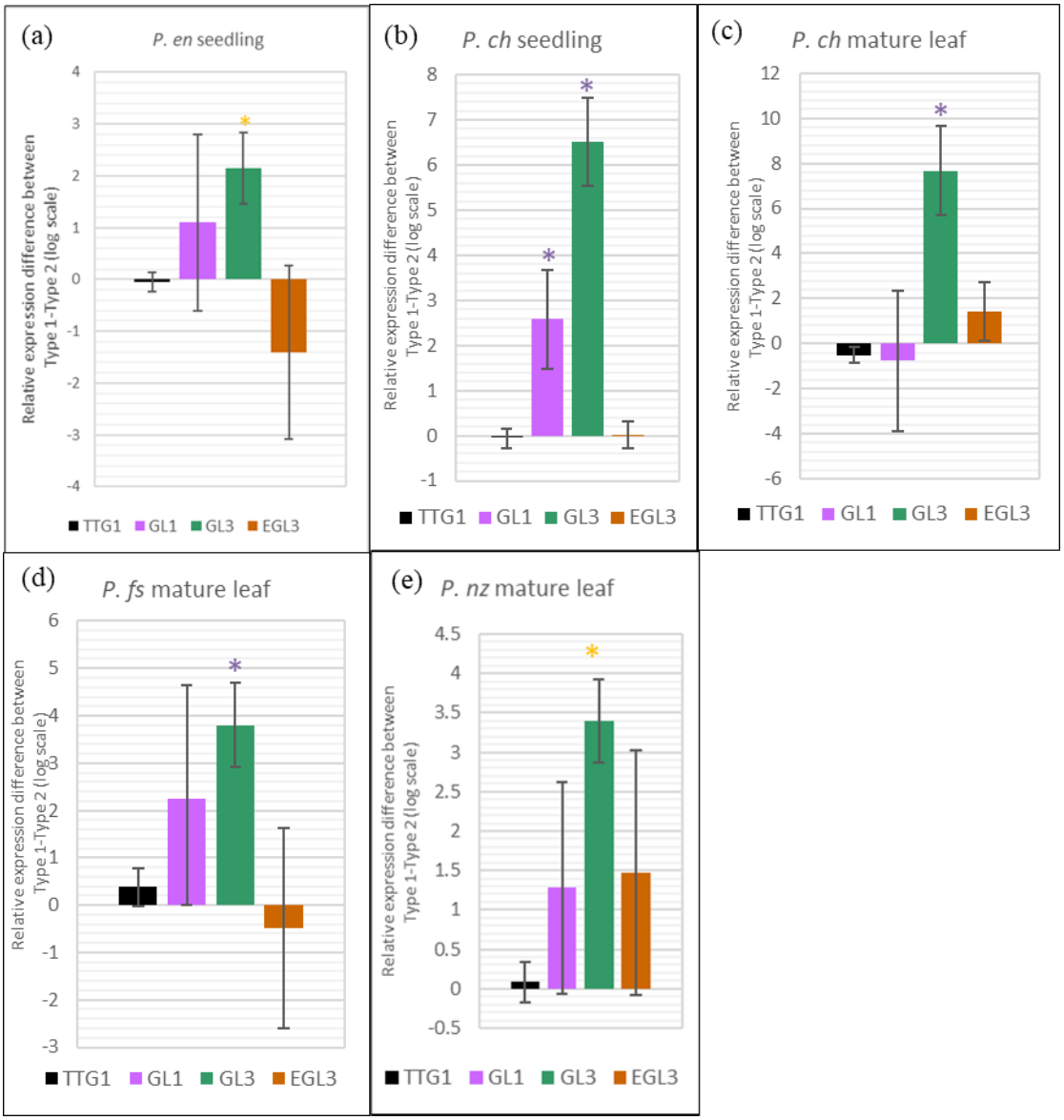


Figure 4.3: Log 2 fold difference in mean expression of Type 1 homeolog (A) to that of the Type 2 homeolog (B) (A-B log scale) in (a) *P. en* (*P. enysii*) seedling (b) *P.ch* (*P. cheesemanii*) seedling (c) *P.ch* (*P. cheesemanii*) mature leaf, (d) *P. fs* (*P. fastigiatum*) mature leaf and (e) *P. nz* (*P. novae-zealandiae*) mature leaf. The error bars represent a 5% to 95% confidence interval of the mean log fold differences. Positive difference shows more expression in Type 1 than in Type 2, and negative difference shows more expression in Type 2 than in Type 1. Significant homeolog expression differences less than the p-value 0.01274 (corrected P-value threshold for multiple testing) are indicated using yellow asterisks, and p-value less than 0.05 (default threshold) are indicated using blue asterisks.

The t-tests evaluate if there is a significant difference in the mean relative gene expression between two groups of data, e.g., between *P. enysii* GOI Type 1 and Type 2, etc. The null hypothesis is that there is no difference between the group means, thus a low p-value

<0.01274 indicates that the null hypothesis should be rejected, indicating that there is a significant difference in gene expression between the two groups (homeologs).

4.4.2.1 Consistently different homeolog expression of *GL3* in all species and developmental stages, a species-specific homeolog expression difference in *GL1*, and a general trend of parental bias

Based on the results obtained (Figure 4.3), a significant differential expression of homeologs was identified. In *P. enysii* seedlings (Table 4.6a and Figure 4.3a), mature leaves of *P. fastigiatum* (Table 4.6d and Figure 4.3d), mature leaves of *P. novae-zealandiae* (Table 4.6e and Figure 4.3e) and mature leaves and seedlings of *P. cheesemanii* (Table 4.6c and Figure 4.3c), *GL3* Type 1 is expressed significantly more than the Type 2 homeolog. In young *P. cheesemanii*, the Type I *GL1* homeolog is significantly expressed more than the Type 2 homeolog (Table 4.6b and Figure 4.3b). Thus, interestingly, in all cases of significant differential expression, it is the Type 1 homeolog that has greater expression relative to the Type 2 homeologs.

Notably, the *GL3* homeolog expression difference is the greatest and the only one that is significant across all species and tissues tested. The values ranged from log 2 to nearly log 8 times greater expression of Type 1and the highest differences were observed in *P. cheesemanii* young and mature plants as log 6.5 times and log 7.7 times (Table 4.6b and Table 4.6c).

At the corrected significance threshold, the only *GL1* significant expression difference was observed in young *P. cheesemanii*, which amounted to a difference of 2.6 times in log scale and equates to a 6-fold linear expression difference in Type 1 compared to Type 2 (Table 4.6b and Figure 4.3b). Notably, the significant difference of *GL1* Type 1 in young *P. cheesemanii* disappears in mature leaves, where both homeologs have comparatively low and similar expression levels (Table 4.6c, Figure 4.3c). Across all species, *TTG1* and *EGL3* show the least differentiation for homeolog expression, whereas only *EGL3* homeologs in *P. novae-zealandiae* have significantly different expression levels (before multiple-test correction).

Interestingly, the relative expression levels of *TTG1* homeologs is similar within each species/tissue examined.

4.4.2.2 Overall biased expression of Type 1 gene copies

Considering significant homeologs expression differences across all species, the results show that the Type 1 homeologs tend to be more highly expressed than Type 2 copies, most notably in *GL3* and *GL1*. This same trend is widely observed across species and genes even where significant differences were not detected. There are a few examples where Type 2 copies are expressed slightly more than Type 1 copies (most notably in *EGL3* in *P. enysii* seedlings and *TTG1* in *P. cheesemanii* mature leaves), but the highest fold change of expression out of these is only 2.6 times (*EGL3* in *P. enysii*).

4.4.2.3 Homeolog expression differences in *P. cheesemanii* seedlings vs. mature leaves

To investigate potential differences in homeolog expression in young seedlings (where trichomes are still being initiated) and fully expanded mature leaves (where trichome initiation has ceased), gene expression was measured from these two tissues in *P. cheesemanii* (this was the only species where sufficient material allowed this comparison). The results are summarised in Table 4.7 and Figure 4.4, which compares expression differences of all Type 1 and Type 2 GOI homeologs.

The largest expression difference between these two tissues was observed in *GL1*, where the Type 1 homeolog was found to be expressed 51 times more in seedlings than in mature leaves (Figure 4.4a and b). The *GL1* Type 2 homeolog was five times more expressed in seedlings compared with mature leaves (Table 4.7a and b), but this difference was only significant prior to multiple-test correction.

No other significant expression differences between the two *P. cheesemanii* tissues were detected following multiple-test correction. The *TTG1* Type 2 homeolog showed greater expression in mature leaves than in seedlings and the *EGL3* Type 1 expression was modestly greater in seedlings than in mature leaves, but the former was not significant following

multiple-test correction, and the latter was just outside significance even before correction. The two copies of *GL3* were each very similarly expressed in both tissue samples.

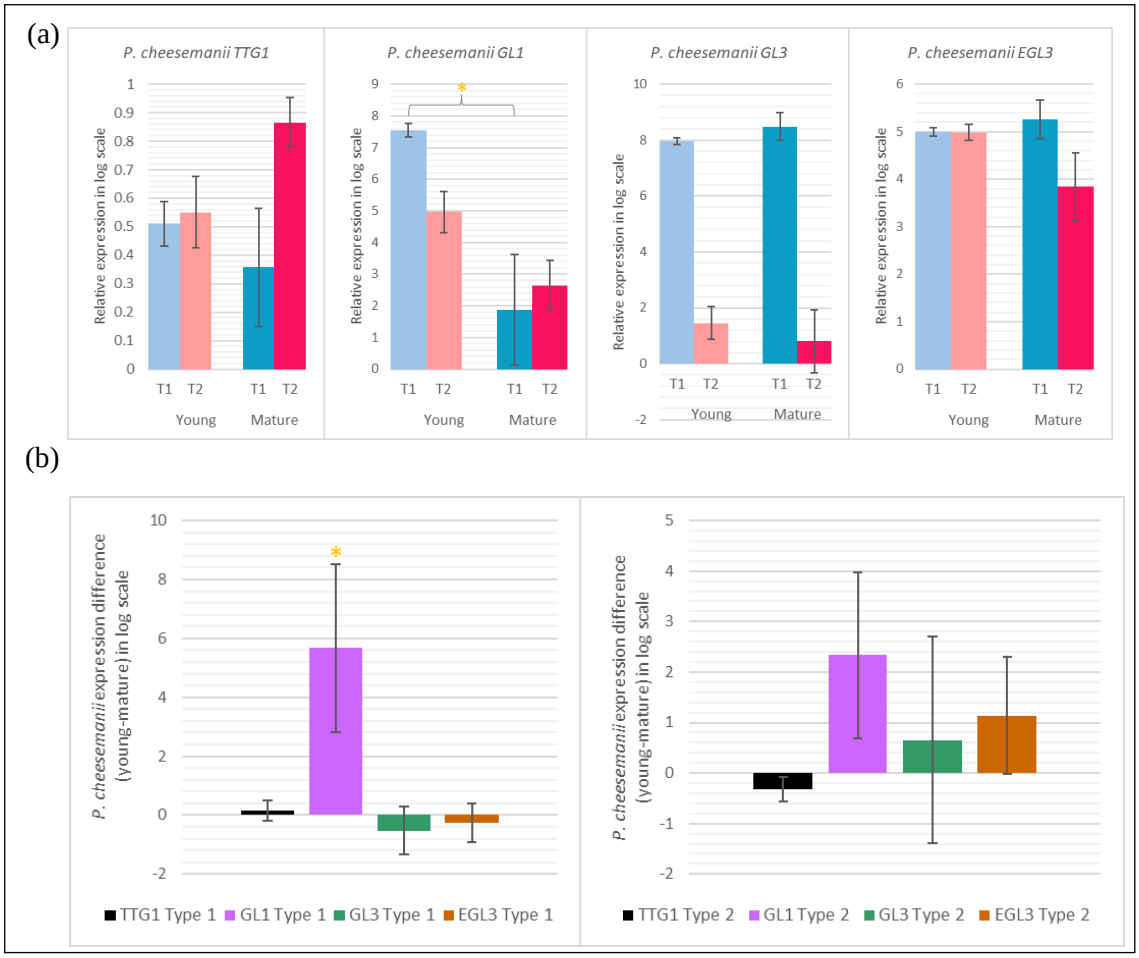


Figure 4.4: (a) Log 2 scale relative expression differences in type-specific homeologs of the genes in young and mature *P. cheesemanii*. (b) Log fold expression difference of Type 1 homeologs and (c) Type 2 homeologs of all genes in young *P. cheesemanii* compared to its mature stage. Positive values exhibit the particular homeolog is expressed more in the young stage while negative values indicate it is expressed more in the mature stage. The error bars represent a 5% to 95% confidence interval. Asterisk marks indicate significant expression differences that are less than the p-value 0.01274 (corrected P-value threshold for multiple testing) are highlighted in yellow, and also the P-values less than the default threshold p-value of 0.05 are highlighted in blue.

Table 4.7: Unpaired 2-tail t-test results between (a) Type 1 homeologs and (b) Type 2 homeologs of all genes (*TTG1*, *GL1*, *GL3*, and *EGL3*) between young *P. cheesemanii* (*P.ch*) and mature *P. cheesemanii*. The first three rows of all tables show pre-processed log Cp measurements, then below them the number of measurements (Count), the average measurement (Mean), the standard deviation from the mean (STDEV), the degrees of freedom (df), the test statistic (t), the probability (p-value), standard error of the mean (SEM) and t-value (t*). The difference in A (young *P.ch*) - B (mature *P.ch*) is shown in the graphs in Figure 4.4. The important 'P-value' row is coloured in grey and further, the significant P-values less than 0.01274 (corrected P-value threshold for multiple testing) are highlighted in yellow, and also the P-values less than default threshold p-value of 0.05 are highlighted in blue.

(a)

	TTG1*_T1 young *P.ch	***TTG1*_T1 mature *P.ch***	***GL1*_T1 young *P.ch***	***GL1*_T1 mature *P.ch***	***GL3*_T1 young *P.ch***	***GL3*_T1 mature *P.ch***	***EGL3*_T1 young *P.ch***	***EGL3*_T1 mature *P.ch***
	0.4792	0.54979	7.73249	0.12055	8.04339	7.91427	4.88473	5.04682
	0.59855	0.38695	7.61713	1.86854	8.02145	8.75839	5.05133	4.99428
	0.45237	0.1373	7.31387	3.62651	7.83598	8.79224	5.05084	5.72605
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.51004	0.35802	7.5545	1.87187	7.96694	8.4883	4.99563	5.25571
STDEV	0.07782	0.20776	0.21622	1.75299	0.11394	0.49741	0.09605	0.40817
df		4.		4.		4.		4.
SD^2		0.02461		1.55986		0.1302		0.08791
t		1.18686		5.57254		1.76961		1.07431
P-Value		0.300953491		0.005081764		0.151509135		0.34317007
Confidence level		0.95		0.95		0.95		0.95
SEM		0.12809		1.01976		0.29462		0.24209
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		0.15202		5.68263		-0.52136		-0.26008
CI start/end (log)	-0.20361	0.50766	2.85133	8.51393	-1.33936	0.29663	-0.93224	0.41207
Fold change		1.11113		51.36209		-1.43531		-1.19755
CI start/end (linear)	-1.15158	1.42174	7.21667	365.55139	-2.53038	1.22828	-1.90823	1.3306

(b)

	TTG1*_T2 young *P.ch	***TTG1*_T2 mature *P.ch***	***GL1*_T2 young *P.ch***	***GL1*_T2 mature *P.ch***	***GL3*_T2 young *P.ch***	***GL3*_T2 mature *P.ch***	***EGL3*_T2 young *P.ch***	***EGL3*_T2 mature *P.ch***
	0.68471	0.90384	5.61657	1.78621	1.64603	0.	4.87682	3.32294
	0.53379	0.92658	4.98482	2.77365	1.93981	0.32372	5.1727	3.55537
	0.43406	0.76765	4.31619	3.36087	0.79412	2.10157	4.90021	4.65035
Count	3.	3.	3.	3.	3.	3.	3.	3.
Mean	0.55085	0.86602	4.97253	2.64024	1.45998	0.80843	4.98324	3.84289
STDEV	0.12619	0.08595	0.65028	0.79576	0.59507	1.13153	0.16449	0.70887
df		4.		4.		4.		4.
SD^2		0.01166		0.52805		0.81723		0.26478
t		3.57532		3.93089		0.88272		2.71421
P-Value		0.023264845		0.017090238		0.427243907		0.053304812
Confidence level		0.95		0.95		0.95		0.95
SEM		0.08815		0.59332		0.73812		0.42014
t*		2.77645		2.77645		2.77645		2.77645
Difference (A-B log scale)		-0.31517		2.33228		0.65155		1.14035
CI start/end (log)	-0.55992	-0.07042	0.68496	3.97961	-1.3978	2.70091	-0.02615	2.30686
Fold change		-1.24416		5.03601		1.57086		2.20435
CI start/end (linear)	-1.47419	-1.05002	1.60765	15.77543	-2.63499	6.5021	-1.01829	4.94804

4.5 DISCUSSION

Following allopolyploidization, there are many changes to genome structure and function that take place. Although this has been well-studied generally in terms of genome-wide expression changes (Bottani et al 2018, Buggs et al 2014, Grover et al 2012, Jenczewski et al 2013), far less has been published in the context of individual genetic pathways. Here we explore the expression of duplicated gene copies (homeologs) in the well-defined *Brassicaceae* trichome initiation pathway in species of the allopolyploid genus *Pachycladon*, coupling results from the previous chapters on isolating homeologs and identifying their relative origins and on molecular evolution of the duplicated pathway. *Pachycladon* was identified as an ideal study system for investigating the fate(s) of duplicated genes in a pathway context because of its evolutionary history, starting with highly diverged progenitor lineages, hybridization/allopolyploidy, and subsequent diversification into multiple species (Joly et al 2009, McBreen & Heenan 2006, Yogeeswaran et al 2011) . In this chapter, we investigate the patterns of expression for homeologs of four genes in the Brassicaceae trichome initiation regulatory complex in four species of *Pachycladon*.

4.5.1 *TTG1* sequence diversity and homeolog expression

No significant differences in *TTG1* homeolog expression were detected after correcting for multiple tests, although subtle, but insignificant differences were observed in mature leaves of *P. fastigiatum* and *P. cheesemanii*. Interestingly, the homeolog expression differences in these two samples were in opposite directions, i.e., in *P. fastigiatum*, it was the Type 1 homeolog that was more highly expressed and in *P. cheesemanii*, it was the Type 2 homeolog. For the other three species/tissue combinations, no appreciable homeolog expression difference was detected.

The DNA polymorphism analysis in Chapter 3 found that the *TTG1* Type 1 homeologs across all *Pachycladon* species examined are slightly more conserved in terms of nucleotide diversity compared to the Type 2 homeolog. Within each homeolog, low nucleotide diversity was found across the coding region (Type 1 = 0.00648 and Type 2 = 0.01053), including the

WD domain and C-terminal end, which are thought to be important in the interaction with bHLHs (Galway et al 1994, Walker et al 1999, Zhang et al 2019a). The Ka/Ks analysis of *TTG1* indicated that purifying selection has been acting on both *TTG1* homeologs. Only one region showed a Ka/Ks ratio greater than one, potentially indicating a region of very local diversifying selection in the two progenitors. Master regulatory genes are sometimes reduced to single copy following duplication (Arrigo & Barker 2012, Edger & Pires 2009, Roth et al 2007), and retained in multiple copies otherwise. The cotton genome has been shown to have retained two functional copies of *TTG1* (Humphries et al 2005). Given the highly pleiotropic nature of *TTG1* and that it forms part of a transcription complex, mutations to either copy that affect structure or function might be expected to have dramatic effects on the fitness of plants that carry those mutations. This might explain the maintenance and low diversity of the two copies.

4.5.2 *GL1* sequence diversity and homeolog expression

Across species, the Type I homeologs of *GL1* tended to show greater expression than Type 2 homeologs; however, the only statistically significant difference detected was in *P. cheesemanii* seedlings. This is in contrast with the results obtained for the *P. cheeesemanii* mature leaves, which showed the least differentiation for *GL1* homeologs of all samples. Indeed, the Type I *GL1* homeolog showed a significant difference in expression in *P. cheesemanii* seedlings vs mature leaves (~50 times higher expression in seedlings vs mature leaves) (Figure 4.4b). Following the same pattern, Type 2 expression is reduced ~five times in mature leaves compared to seedlings (Figure 4.4c). Hence, the relative expression of both *GL1* homeologs is considerably greater in seedlings vs mature leaves in *P. cheesemanii*. This result is consistent with the known timing of peak *GL1* expression in young leaf primordia in *Arabidopsis thaliana* (Szymanski et al 1998). As the seedling samples included the shoot apical meristem, even the youngest leaf primordia with incipient trichomes were captured compared with mature leaves, which are no longer initiating trichomes.

Similar to the *TTG1* result, in the previous chapter we observed that the *GL1* Type 2 homeolog has slightly more (0.013) nucleotide diversity than Type 1 (0.009), including within the functionally important R2R3 and transcriptional activation domains. Ka/Ks analysis revealed that both *GL1* homeologs are under purifying selection across most of the coding sequence. Compared with *TTG1*, *GL1* has a much narrower range of phenotypes, mostly known to be involved in trichome initiation and stomata patterning (Tsuji & Coe 2013). Also, it has been shown in *A. thaliana* that natural accessions harbor considerable variation for *GL1* that has both qualitative and quantitative effects on trichome density (Bloomer et al 2012). Based on this, perhaps natural selection on *Pachycladon* species has been more tolerant of expression changes in *GL1*.

Among the more interesting findings in the sequence analyses was a five-base pair deletion in exon 1 of the *P. fastigiatum GL1* Type 2 homeolog that results in a translation frameshift and an early stop codon. This translates into a severely truncated polypeptide that terminates within the important R2 MYB domain. With the loss of 75% of the protein, this product clearly cannot function at its normal capacity. This is a significant finding as *P. fastigiatum* is glabrous and it is clear from work in *A. thaliana* and other groups (Bloomer et al 2012, Hahn et al 2018, Meng et al 2023, Muto & Matsumoto 2022) that a nonfunctional *GL1* results in glabrous plants. Of course, *Pachycladon* species have two copies of *GL1* and the Type 1 copy in *P. fastigiatum* appears to be intact. Interestingly, both copies are still expressed in *P. fastigiatum*; the Type 1 homeolog is expressed nearly five times more than the Type 2 copy, but it is not a statistically significant difference. So, why is *P. fastigiatum* glabrous if it still has a functional copy of *GL1* being expressed? One possibility is that the truncated Type 2 GL1 protein is somehow interfering with the normal assembly or functioning of the activation complex.

4.5.3 *GL3* sequence diversity and homeolog expression

Interestingly, our results exhibited significantly greater expression of the *GL3* Type 1 homeolog than the Type 2 homeolog for all *Pachycladon* species examined and for both

seedlings and mature leaves. This trend is similar to what was seen with *GL1* but is much stronger for *GL3*. While *GL3* is known to be pleiotropic like *TTG1*, it also demonstrates some functional overlap with *EGL3* as described in *A. thaliana* (Patra et al 2013, Zhang et al 2003). As such, this may have provided some buffering and allowed for greater tolerance to expression changes. Alternatively, sequence differences between Type 1 and Type 2 homeologs of *GL1* and *GL3* may translate into protein differences that limit the known protein-protein interactions between the two. That is, the interactions between Type 1 GL1 and Type 2 GL3 (and the reciprocal combination) may be less functional (and thus confer lower fitness) compared with the original combinations (e.g., Type 1 GL1 and Type 1 GL3). In this scenario, an expression bias in one gene would likely drive the evolution of a corresponding expression change in the other gene. For example, if mutations arose that led to the biased expression of Type 1 *GL3*, then this might lead to natural selection favoring mutations that led to the biased expression of Type 1 *GL1*. While we might expect such a scenario to also lead to biased expression of Type 1 *TTG1*, as TTG1 is also known to physically interact with GL3 (and EGL3), the interaction domains of the homeologs of GL3 and TTG1 may not have inhibitory mutations. Unfortunately, complete sequence sets were not acquired for *GL3*, so direct comparisons between expression profiles and genetic variation among Types cannot yet be made. However, the fact that the same significant expression bias is observed for *GL3* across all species and tissue types investigated, suggests that the change could have arisen in the initial *Pachycladon* lineage prior to species diversification. A similar type of scenario describing the reorganization of biological pathways after whole genome duplication has been discussed by De Smet and Van de Peer (2012) in the means of gene dosage balance.

Our *GL3* sequences did not permit for complete sequence analysis due to lack of complete coding sequences, but the amount of sequences we were able to find revealed variation with a few conspicuous mutations, which can be coupled with the expression pattern of the homeologs described above. The lower expression of *GL3* Type 2 copies may be related to the frameshift mutations observed in *P. cheesemani*, *P.enysii*, and *P. fastigiatum* (*P. novae-*

zelandiae Type 2 sequence was not complete enough to identify any obvious detrimental mutations). In addition, *P. cheesemanii GL3* Type 2 sequences showed a mutation in the start codon, meaning that the transcript is not likely to be translated properly. The expression difference between Types 1 and 2 of *GL3* in *P. cheesemanii* seedlings and mature leaves represent the greatest differences observed here. If mutations accumulated in *GL3* Type 2 sequences that prevented proper translation and protein function, then, as with the Type 2 homeolog of *GL1* in *P. fastigiatum*, it is easy to imagine natural selection favoring subsequent mutations that would have reduced its expression.

4.5.4 *EGL3* sequence diversity and homeolog expression

EGL3 showed considerable diversity for expression profiles. Although none of the comparisons for expression between Type 1 and 2 were statistically significant for *EGL3*, there are clear patterns observed that include higher expression of Type 1 in *P. novae-zelandiae* and *P. cheesemanii* mature leaves, higher expression of Type 2 in *P. enysii* mature leaves, and very similar expression levels of the two homeologs in *P. fastigiatum* mature leaves and *P. cheesemanii* seedlings. Like *GL3*, *EGL3* is an interesting gene as it is both pleiotropic and overlaps in function with *GL3* (Morohashi & Grotewold 2009, Zhang et al 2003). This particular combination may underlie its lability in expression patterns observed across species and tissue types. In contrast with *GL3*, it seems more likely that new mutations and selection pressures have arisen in *EGL3* independently in each lineage, leading to the highly variable patterns observed. Unfortunately, insufficient sequence data were obtained to make direct comparisons between expression patterns and nucleotide diversity patterns.

4.5.5 Comparison between mature leaf tissue and seedling gene expression

Since similar RNA concentrations were used for *P. cheesemanii* seedlings and mature leaves, it was feasible to compare the expression of gene homeologs between the two growth stages. From all the homeologs, only *GL1* Type 1 was found to be expressed significantly differently in young seedlings compared to mature leaves, while Type 2 was also expressed

more in seedlings but only showed a significant difference prior to multi-test correction. So, both homeologs are more highly expressed in the seedling compared to the mature leaf.

TTG1 Type 1 was slightly more expressed in the seedling compared to the mature leaf. Type 2 was expressed more in mature leaves which is significant only before multi-test correction. *GL3* Type 1 and Type 2 are similarly expressed in both seedlings and mature leaves. *EGL3* Type 1 was expressed more in mature leaves while the Type 2 copy had more expression in the seedling.

Interestingly, in chamber-grown *P. cheesemanii* seedlings, we noticed that the first leaves to emerge were glabrous and later leaves developed branched trichomes. Generally, *GL1* is a decisive factor in regulating trichome initiation and loss of function mutations in *GL1* lead to a glabrous phenotype in *Arabidopsis* lab strains and natural populations (Bloomer et al 2012, Kawagoe et al 2011, Sato & Kudoh 2015). However, the glabrous phenotype of the first true leaves of *P. cheesemanii* is not based on coding mutations in *GL1*, because neither *GL1* homeolog exhibited any detrimental mutations and were expressed (Type 1 copy significantly expressed) in the seedlings. Moreover, at least one copy of other important genes (*TTG1*, *EGL3*, and *GL3*) which are required for the trichome initiation complex were/was expressed in seedlings. Therefore, the glabrous phase of young leaves could not be explained by gene expression patterns we have already studied. Although environmental cues and other upstream and downstream factors which we have not considered in this thesis could have an effect; e.g. GLABROUS INFLORESCENCE STEMS (GIS) transcription factor (Gan et al 2006) regulates the expression of *GL1*.

4.5.6 A role for pleiotropy in homeolog expression and retention

TTG1 homeologs showed more or less similar expression across the species and growth stages. In *A. thaliana*, *TTG1* is highly pleiotropic and controls multiple important developmental pathways such as trichome and root hair differentiation, flavonoid biosynthesis, and seed coat mucilage deposition (Gonzalez et al 2009, Lu et al 2009, Zhang et al 2003). Hence, it seems the retention and expression of *TTG1* homeologs are exceptional compared to

those of the other genes, tending to keep both the duplicated copies in the polyploid genome and express them over a long evolutionary period. Having originated long ago (Doroshkov et al 2019), *TTG1* seems to be maintained without much divergence. The Ka/Ks analyses performed in Chapter 3 revealed that the two ancestral lineages that hybridized to form *Pachycladon* possessed *TTG1* copies that were very similar in terms of amino acid sequence.

Pleiotropy is thought to be one of the mechanisms preserving the redundancy of young duplicated genes (Krakauer & Nowak 1999). For instance, in tetraploid cotton, four copies of the *TTG1* gene have been retained and two of them (*GhTTG1* and *GhTTG3*) were able to restore trichome formation in *ttg1* mutant *Arabidopsis* plants (Humphries et al 2005). These two gene copies were suggested to have pleiotropic roles controlling cotton fiber initiation and production of anthocyanin pigments. In the *Brassicaceae*, two *TTG1* genes were found in *Matthiola incana*, and one was shown to have pleiotropic roles including trichome initiation while the other *TTG1*-like gene's function is not known (Dressel & Hemleben 2009). Moreover, in some *Brassica* species such as *B. juncea*, *B. carinata* (Yan et al 2014) and *B. oleracea* (Nayidu et al 2014), two *TTG1* copies were found and in *B. napus*, six copies were found originating from its three parental subgenomes (Nayidu et al 2014). However, in other cases, *TTG1* has been reduced to a single copy, as in *A. thaliana*, *Malcolmia flexuosa* (Dressel & Hemleben 2009) and *B. villosa* (Nayidu et al 2014).

4.5.7 A general finding of biased expression of Type 1 homeologs over Type 2

When examining the expression patterns of the homeologs of the four genes, a general finding supported our hypothesis by expressing Type 1 homeologs more than Type 2 homeologs in most cases, including all significant expression differences (Figure 4.3). In a few instances, the Type 2 homeolog was slightly expressed than the Type 1 homeolog; i. e; *TTG1* and *EGL3* in *P. enysii* seedlings, *TTG1* in *P. cheesemanii* seedling, *TTG1* and *GL1* in *P. cheesemanii* mature leaf and *EGL3* in *P.fastigiatum* mature leaf.

Based on previous studies (Hendriks et al 2023, Huang et al 2015, Joly et al 2009), it was discussed in Chapter 2 that Type 2 homeologs potentially originated from the maternal

progenitor, so the Type 1 homeolog originated from the paternal progenitor, the latter of which exhibited biased expression for all genes we investigated in the four *Pachycladon* species. This phenomenon, that homeologs are expressed at unequal levels is referred to as homeolog expression bias (Grover et al 2012). There are many examples from the literature of one of the subgenomes having expression bias in allopolyploids, such as in maize (Schnable & Freeling 2011), cotton (Yoo et al 2013, Zheng et al 2016a), and wheat (Kumar et al 2022) where it is more common that there is a maternal bias (Sehrish et al 2015, Videvall et al 2015, Wang et al 2017, Wu et al 2016), although paternal bias has also been observed (Guo et al 2006, Moreno-Romero et al 2019, Zhang et al 2021).

As discussed in Chapter 3, Type 1 homeologs seem to be under stronger purifying selection than Type 2 homeologs (observed as less genetic variation in Type 1 compared with Type 2). This is consistent with other studies that have reported such biased subgenome expression occurring with enhanced purifying selection (Hughes et al 2014, Pophaly & Tellier 2015). On the other hand, biased fractionation is also promoted in relatively less expressed homeologs (Cheng et al 2012, Freeling et al 2015). Based on our results, Type 2 homeologs are relatively less expressed and possess more evidence for fractionation; e. g. *P. fastigiatum GL1* Type 2 and *P. cheesemanii GL3* Type 2.

4.5.8 Evolution of relative expression of *TTG1, GL1, GL3,* and *EGL3* in the *Pachycladon* trichome initiation complex is suggested to be maintained by combinatorial effects of subgenome dominance, biased fractionation, and relative gene dosage balance

The observations made here in gene expression differences in the homeologs of interest are likely underpinned by multiple mechanisms, e.g. relative dosage, absolute dosage, subgenome dominance, and biased fractionation. These mechanisms are often not easily separated, but instead, they may overlap, interact, shift, and vary over time (Conant et al 2014).

Generally, soon after the whole genome duplication that took place in *Pachycladon*, all the duplicate gene pairs may have been expressed redundantly as happens in some nascent polyploid lineages (Buggs et al 2011). However, a common occurrence is methylation shock

which can trigger biased expression straightaway (Zhang et al 2018). While redundancy is often observed in the duplicates from the most recent WGD events in *Arabidopsis*, largescale studies report that redundancy seems to have been lost for the older *A. thaliana* paralogs and hence homeolog redundancy might just represent a transient state (De Smet & Van de Peer 2012). Thus, in the context of the *Pachycladon* trichome initiation complex and pathway, both copies from all genes may have been functional at the start of the lineage and continued to function, similarly regulated by the relative dosage balance of the component genes. Eventually though, a new balance must be established based on mutations and potential methylation changes (De Smet & Van de Peer 2012). Another possibility is that the two homeologs might have had unequal expressions from the beginning (passed on from the progenitors) and that pattern was not changed and carried on after the hybridization and polyploidization event(s).

Here we have observed a general (though not exclusive) bias toward the expression of Type 1 homeologs. This matches a general finding of likely stronger purifying selection acting on the Type 1. Likewise, when fractionating the homeologous copies, a biased fractionation towards the Type 2 subgenome has been seen whereas Type 1 homeologs tend to retain and dominate the gene expression. With time, a single Type 1 subgenome may contribute to the majority of the phenotype as reviewed in Alger and Edger (2020).

Some of the homeologs in our dataset were confirmed to be pseudogenes in the sequence analyses (*P. fastigiatum GL1* Type 2) but are still expressed. Pseudogenes are generally identified based on their similarity to annotated genes and the presence of detrimental mutations such as frameshifts in protein-coding genes and premature stop codons that result in loss of function (Vanin 1985). But, some pseudogenes may still be functional as truncated proteins or as RNA (Siena et al 2016), though often with different functions.

4.6 CONCLUSION

Some elements of the *Pachycladon* trichome initiation complex show differential homeolog expression after ~1.6 million years of evolution and we suggest multiple molecular mechanisms underlie the present expression status in the group. Notably, expression bias of Type 1 homeologs over Type 2 homeologs was observed across species, generally significant for *GL3* and *GL1*. For *TTG1* both copies showed very similar expression across genes and species perhaps due to its strong pleiotropic nature. The *EGL3* homeolog expression profiles were quite variable, but significant differences were not observed. Less expressed *GL1* and *GL3* Type 2 copies showed some detrimental mutations. Overall, the duplicated trichome initiation pathway seems to be still resolving itself in *Pachycladon* species, yet quite independently within individual species.

REFERENCES

IDT OligoAnalyzer, pp. Integrated DNA Technologies.

IDT UNAFold, pp. Integrated DNA Technologies.

Alger, E. I., and P. P. Edger, 2020 One subgenome to rule them all: underlying mechanisms of subgenome dominance. Current opinion in plant biology 54**:** 108-113.

Arezi, B., W. Xing, J. A. Sorge and H. H. Hogrefe, 2003 Amplification efficiency of thermostable DNA polymerases. Anal Biochem 321**:** 226-235.

Arrigo, N., and M. S. Barker, 2012 Rarely successful polyploids and their legacy in plant genomes. Current opinion in plant biology 15**:** 140-146.

Balkunde, R., M. Pesch and M. Hülskamp, 2010 Trichome patterning in Arabidopsis thaliana: from genetic to molecular models. Current topics in developmental biology 91**:** 299-321.

Bekaert, M., P. P. Edger, J. C. Pires and G. C. Conant, 2011 Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. The Plant Cell 23**:** 1719-1728.

Birchler, J. A., U. Bhadra, M. P. Bhadra and D. L. Auger, 2001 Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol 234**:** 275-288.

Birchler, J. A., and R. A. Veitia, 2010 The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytologist 186**:** 54-62.

Birchler, J. A., H. Yao and S. Chudalayandi, 2007 Biological consequences of dosage dependent gene regulatory systems. Biochim Biophys Acta 1769**:** 422-428.

Bird, K. A., R. VanBuren, J. R. Puzey and P. P. Edger, 2018 The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytologist 220**:** 87-93.

Blanc, G., and K. H. Wolfe, 2004 Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16**:** 1679-1691.

Bloomer, R. H., T. E. Juenger and V. V. Symonds, 2012 Natural variation in GL1 and its effects on trichome density in Arabidopsis thaliana. Molecular Ecology 21**:** 3501-3515.

Bomblies, K., and J. F. Doebley, 2006 Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes zfl1 and zfl2 on traits under selection during maize domestication. Genetics 172**:** 519-531.

Bottani, S., N. R. Zabet, J. F. Wendel and R. A. Veitia, 2018 Gene expression dominance in allopolyploids: hypotheses and models. Trends in Plant Science 23**:** 393-402.

Bouyer, D., F. Geier, F. Kragler, A. Schnittger, M. Pesch *et al.*, 2008 Two-dimensional patterning by a trapping/depletion mechanism: the role of TTG1 and GL3 in Arabidopsis trichome formation. PLoS biology 6**:** e141.

Buggs, R. J., J. F. Wendel, J. J. Doyle, D. E. Soltis, P. S. Soltis *et al.*, 2014 The legacy of diploid progenitors in allopolyploid gene expression patterns. Philosophical Transactions of the Royal Society B: Biological Sciences 369**:** 20130354.

Buggs, R. J., L. Zhang, N. Miles, J. A. Tate, L. Gao *et al.*, 2011 Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol 21**:** 551-556.

Burns, R., T. Mandáková, J. Gunis, L. M. Soto-Jiménez, C. Liu *et al.*, 2021 Gradual evolution of allopolyploidy in Arabidopsis suecica. Nature ecology & evolution 5**:** 1367-1381.

Bustin, S., and J. Huggett, 2017 qPCR primer design revisited. Biomolecular Detection and Quantification 14**:** 19-28.

Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett *et al.*, 2009 The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55**:** 611-622.

Chen, Z. J., 2007 Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annual Review of Plant Biology 58**:** 377-406.

Chen, Z. J., 2010 Molecular mechanisms of polyploidy and hybrid vigor. Trends in plant science 15**:** 57-71.

Cheng, F., C. Sun, J. Wu, J. Schnable, M. R. Woodhouse *et al.*, 2016 Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytologist 211**:** 288-299.

Cheng, F., J. Wu, L. Fang, S. Sun, B. Liu *et al.*, 2012 Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. PloS one 7**:** e36442.

Colle, M., C. P. Leisner, C. M. Wai, S. Ou, K. A. Bird *et al.*, 2019 Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. GigaScience 8**:** giz012.

Conant, G. C., J. A. Birchler and J. C. Pires, 2014 Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Current Opinion in Plant Biology 19**:** 91-98.

Czechowski, T., M. Stitt, T. Altmann, M. K. Udvardi and W. R. Scheible, 2005 Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. Plant Physiol 139**:** 5-17.

De Smet, R., E. Sabaghian, Z. Li, Y. Saeys and Y. Van de Peer, 2017 Coordinated Functional Divergence of Genes after Genome Duplication in Arabidopsis thaliana The Plant Cell 29**:** 2786-2800.

De Smet, R., and Y. Van de Peer, 2012 Redundancy and rewiring of genetic networks following genome-wide duplication events. Current opinion in plant biology 15**:** 168-176.

Dekkers, B. J. W., L. Willems, G. W. Bassel, R. P. van Bolderen-Veldkamp, W. Ligterink *et al.*, 2011 Identification of Reference Genes for RT–qPCR Expression Analysis in Arabidopsis and Tomato Seeds. Plant and Cell Physiology 53**:** 28-37.

Doroshkov, A. V., D. K. Konstantinov, D. A. Afonnikov and K. V. Gunbin, 2019 The evolution of gene regulatory networks controlling Arabidopsis thaliana L. trichome development. BMC Plant Biology 19**:** 53.

Dressel, A., and V. Hemleben, 2009 Transparent Testa Glabra 1 (TTG1) and TTG1-like genes in Matthiola incana R. Br. and related Brassicaceae and mutation in the WD-40 motif. Plant Biology 11**:** 204-212.

Edger, P. P., and J. C. Pires, 2009 Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research 17**:** 699-717.

Edger, P. P., T. J. Poorten, R. VanBuren, M. A. Hardigan, M. Colle *et al.*, 2019 Origin and evolution of the octoploid strawberry genome. Nature Genetics 51**:** 541-547.

Expósito-Rodríguez, M., A. A. Borges, A. Borges-Pérez and J. A. Pérez, 2008 Selection of internal control genes for quantitative real-time RT-PCR studies during tomato development process. BMC plant biology 8**:** 131-131.

Flagel, L., J. Udall, D. Nettleton and J. Wendel, 2008 Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. BMC biology 6**:** 1-9.

Flatschacher, D., V. Speckbacher and S. Zeilinger, 2022 qRAT: an R-based stand-alone application for relative expression analysis of RT-qPCR data. BMC Bioinformatics 23**:** 286.

Freeling, M., M. J. Scanlon and J. E. Fowler, 2015 Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. Curr Opin Genet Dev 35**:** 110-118.

Galway, M. E., J. D. Masucci, A. M. Lloyd, V. Walbot, R. W. Davis *et al.*, 1994 The TTG gene is required to specify epidermal cell fate and cell patterning in the Arabidopsis root. Developmental biology 166**:** 740-754.

Gan, Y., R. Kumimoto, C. Liu, O. Ratcliffe, H. Yu *et al.*, 2006 GLABROUS INFLORESCENCE STEMS modulates the regulation by gibberellins of epidermal differentiation and shoot maturation in Arabidopsis. The Plant Cell 18**:** 1383-1395.

Garsmeur, O., J. C. Schnable, A. Almeida, C. Jourda, A. D'Hont *et al.*, 2014 Two evolutionarily distinct classes of paleopolyploidy. Molecular biology and evolution 31**:** 448-454.

Geneious 9.1.8, 2017, pp.

Glover, B. J., S. Bunnewell and C. Martin, 2004 Convergent evolution within the genus Solanum: the specialised anther cone develops through alternative pathways. Gene 331**:** 1-7.

Glover, B. J., M. Perez-Rodriguez and C. Martin, 1998 Development of several epidermal cell types can be specified by the same MYB-related plant transcription factor. Development 125**:** 3497-3508.

Gonzalez, A., J. Mendenhall, Y. Huo and A. Lloyd, 2009 TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. Developmental biology 325**:** 412-421.

Grover, C., J. Gallagher, E. Szadkowski, M. Yoo, L. Flagel *et al.*, 2012 Homoeolog expression bias and expression level dominance in allopolyploids. New Phytologist 196**:** 966-971.

Guo, M., M. A. Rupe, X. Yang, O. Crasta, C. Zinselmeier *et al.*, 2006 Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. Theoretical and Applied Genetics 113**:** 831-845.

Ha, M., W. H. Li and Z. J. Chen, 2007 External factors accelerate expression divergence between duplicate genes. Trends Genet 23**:** 162-166.

Hahn, F., M. Eisenhut, O. Mantegazza and A. P. Weber, 2018 Homology-directed repair of a defective glabrous gene in Arabidopsis with Cas9-based gene targeting. Frontiers in Plant Science 9**:** 424.

Hellemans, J., and J. Vandesompele, 2011 qPCR data analysis-unlocking the secret to successful results. PCR troubleshooting and optimization: the essential guide 1**:** 13.

Hendriks, K. P., C. Kiefer, I. A. Al-Shehbaz, C. D. Bailey, A. H. van Huysduynen *et al.*, 2023 Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. Current Biology.

Huang, C.-H., R. Sun, Y. Hu, L. Zeng, N. Zhang *et al.*, 2015 Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. Molecular Biology and Evolution 33**:** 394-412.

Hughes, T. E., J. A. Langdale and S. Kelly, 2014 The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome research 24**:** 1348-1355.

Humphries, J. A., A. R. Walker, J. N. Timmis and S. J. Orford, 2005 Two WD-repeat genes from cotton are functional homologues of the Arabidopsis thaliana TRANSPARENT TESTA GLABRA1 (TTG1) gene. Plant molecular biology 57**:** 67-81.

Ishida, T., S. Hattori, R. Sano, K. Inoue, Y. Shirano *et al.*, 2007 Arabidopsis TRANSPARENT TESTA GLABRA2 is directly regulated by R2R3 MYB transcription factors and is involved in regulation of GLABRA2 transcription in epidermal differentiation. Plant Cell 19**:** 2531-2543.

Jenczewski, E., A.-M. Chèvre and K. Alix, 2013 Chromosomal and gene expression changes in Brassica allopolyploids. Polyploid and hybrid genomics**:** 171-186.

Jin, Y., F. Liu, W. Huang, Q. Sun and X. Huang, 2019 Identification of reliable reference genes for qRT-PCR in the ephemeral plant Arabidopsis pumila based on full-length transcriptome data. Scientific Reports 9**:** 8408.

Joly, S., P. B. Heenan and P. J. Lockhart, 2009 A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. Mol Phylogenet Evol 51**:** 365-372.

Kawagoe, T., K. K. Shimizu, T. Kakutani and H. Kudoh, 2011 Coexistence of trichome variation in a natural plant population: a combined study using ecological and candidate gene approaches. PLoS One 6**:** e22184.

Kirik, V., M. M. Lee, K. Wester, U. Herrmann, Z. Zheng *et al.*, 2005 Functional diversification of MYB23 and GL1 genes in trichome morphogenesis and initiation. Development 132**:** 1749-1749.

Kirik, V., M. Simon, M. Huelskamp and J. Schiefelbein, 2004a The ENHANCER OF TRY AND CPC1 gene acts redundantly with TRIPTYCHON and CAPRICE in trichome and root hair cell patterning in Arabidopsis. Dev Biol 268**:** 506-513.

Kirik, V., M. Simon, K. Wester, J. Schiefelbein and M. Hulskamp, 2004b ENHANCER of TRY and CPC 2 (ETC2) reveals redundancy in the region-specific control of trichome development of Arabidopsis. Plant Mol Biol 55**:** 389-398.

Kliebenstein, D. J., 2008 A role for gene duplication and natural variation of gene expression in the evolution of metabolism. PloS one 3**:** e1838.

Kontanis, E. J., and F. A. Reed, 2006 Evaluation of real-time PCR amplification efficiencies to detect PCR inhibitors. Journal of forensic sciences 51**:** 795-804.

Krakauer, D. C., and M. A. Nowak, 1999 Evolutionary preservation of redundant duplicated genes. Seminars in Cell & Developmental Biology 10**:** 555-559.

Kumar, A., S. Kumar, K. Venkatesh, N. K. Singh, P. K. Mandal *et al.*, 2022 Physio-molecular traits of contrasting bread wheat genotypes associated with 15N influx exhibiting homeolog expression bias in nitrate transporter genes under different external nitrate concentrations. Planta 255**:** 104.

Larkin, J. C., D. G. Oppenheimer, S. Pollock and M. D. Marks, 1993 Arabidopsis GLABROUS1 Gene Requires Downstream Sequences for Function. Plant Cell 5**:** 1739-1748.

Larkin, J. C., N. Young, M. Prigge and M. D. Marks, 1996 The control of trichome spacing and number in Arabidopsis. Development 122**:** 997-1005.

Lemos, B., L. O. Araripe, P. Fontanillas and D. L. Hartl, 2008 Dominance and the evolutionary accumulation of cis-and trans-effects on gene expression. Proceedings of the National Academy of Sciences 105**:** 14471-14476.

Levine, M. T., A. K. Holloway, U. Arshad and D. J. Begun, 2007 Pervasive and largely lineage-specific adaptive protein evolution in the dosage compensation complex of Drosophila melanogaster. Genetics 177**:** 1959-1962.

Li, Q., X. Qiao, H. Yin, Y. Zhou, H. Dong *et al.*, 2019 Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). Horticulture research 6.

Li, S. F., P. J. Allen, R. S. Napoli, R. G. Browne, H. Pham *et al.*, 2020 MYB–bHLH–TTG1 regulates Arabidopsis seed coat biosynthesis pathways directly and indirectly via multiple tiers of transcription factors. Plant and Cell Physiology 61**:** 1005-1018.

Liang, Z., and J. C. Schnable, 2018 Functional divergence between subgenomes and gene pairs after whole genome duplications. Molecular plant 11**:** 388-397.

Liu, X.-L., F.-F. Jiang, Z.-W. Wang, X.-Y. Li, Z. Li *et al.*, 2017 Wider geographic distribution and higher diversity of hexaploids than tetraploids in Carassius species complex reveal recurrent polyploidy effects on adaptive evolution. Scientific Reports 7**:** 5395.

Lloyd, A. M., V. Walbot and R. W. Davis, 1992 Arabidopsis and Nicotiana anthocyanin production activated by maize regulators R and C1. Science 258**:** 1773-1775.

Lu, J., J.-N. Li, B. Lei, S.-G. Wang and Y.-R. Chai, 2009 Molecular cloning and characterization of twoBrassica napus TTG1 genes reveal genus-specific nucleotide preference, extreme protein-level conservation and fast divergence of organ-specificity. Genes & Genomics 31**:** 129-142.

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

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics 154**:** 459-473.

Lynch, M., M. O'Hely, B. Walsh and A. Force, 2001 The probability of preservation of a newly arisen gene duplicate. Genetics 159**:** 1789-1804.

Madlung, A., 2013 Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity 110**:** 99-104.

Martin, C., K. Bhatt, K. Baumann, H. Jin, S. Zachgo *et al.*, 2002 The mechanics of cell fate determination in petals. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 357**:** 809-813.

McBreen, K., and P. B. Heenan, 2006 Phylogenetic relationships of Pachycladon (Brassicaceae) species based on three nuclear and two chloroplast DNA markers. New Zealand Journal of Botany 44**:** 377-386.

Meng, Y., X. Lyu, J. Liu, W. Gao, Y. Ma *et al.*, 2023 Structural variation of GL1 gene determines the trichome formation in Brassica juncea. Theoretical and Applied Genetics 136**:** 75.

Moreno-Romero, J., G. Del Toro-De León, V. K. Yadav, J. Santos-González and C. Köhler, 2019 Epigenetic signatures associated with imprinted paternally expressed genes in the Arabidopsis endosperm. Genome biology 20**:** 1-11.

Morohashi, K., and E. Grotewold, 2009 A Systems Approach Reveals Regulatory Circuitry for Arabidopsis Trichome Initiation by the GL3 and GL1 Selectors. PLOS Genetics 5**:** e1000396.

Morohashi, K., M. Zhao, M. Yang, B. Read, A. Lloyd *et al.*, 2007 Participation of the Arabidopsis bHLH Factor GL3 in Trichome Initiation Regulatory Events. Plant Physiology 145**:** 736-746.

Muller, H. J., 1925 Why Polyploidy is Rarer in Animals Than in Plants. The American Naturalist 59**:** 346-353.

MultiD_Analyses_AB, 2019, pp.

Muto, N., and T. Matsumoto, 2022 CRISPR/Cas9-mediated genome editing of RsGL1a and RsGL1b in radish (Raphanus sativus L.). Frontiers in Plant Science 13**:** 951660.

Nayidu, N. K., S. Kagale, A. Taheri, T. S. Withana-Gamage, I. A. Parkin *et al.*, 2014 Comparison of five major trichome regulatory genes in Brassica villosa with orthologues within the Brassicaceae. PLoS One 9**:** e95877.

Negrón-Ortiz, V., 2007 Chromosome numbers, nuclear DNA content, and polyploidy in Consolea (Cactaceae), an endemic cactus of the Caribbean Islands. American Journal of Botany 94**:** 1360-1370.

Noda, K.-i., B. J. Glover, P. Linstead and C. Martin, 1994 Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor. Nature 369**:** 661-664.

Nougué, O., J. Corbi, S. G. Ball, D. Manicacci and M. I. Tenaillon, 2014 Molecular evolution accompanying functional divergence of duplicated genes along the plant starch biosynthesis pathway. BMC evolutionary biology 14**:** 1-16.

Owczarzy, R., A. V. Tataurov, Y. Wu, J. A. Manthey, K. A. McQuisten *et al.*, 2008 IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res 36**:** W163-169.

Panchy, N., M. Lehti-Shiu and S.-H. Shiu, 2016 Evolution of Gene Duplication in Plants. Plant Physiology 171**:** 2294-2316.

Patra, B., S. Pattanaik and L. Yuan, 2013 Ubiquitin protein ligase 3 mediates the proteasomal degradation of GLABROUS 3 and ENHANCER OF GLABROUS 3, regulators of trichome development and flavonoid biosynthesis in Arabidopsis. The Plant Journal 74**:** 435-447.

Pattanaik, S., B. Patra, S. K. Singh and L. Yuan, 2014 An overview of the gene regulatory network controlling trichome development in the model plant, Arabidopsis. Frontiers in plant science 5**:** 259.

Payne, C. T., F. Zhang and A. M. Lloyd, 2000 GL3 Encodes a bHLH Protein That Regulates Trichome Development in Arabidopsis Through Interaction With GL1 and TTG1. Genetics 156**:** 1349-1362.

Payne, T., J. Clement, D. Arnold and A. Lloyd, 1999 Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in Nicotiana tabacum. Development 126**:** 671-682.

Pfeifer, M., K. G. Kugler, S. R. Sandve, B. Zhan, H. Rudi *et al.*, 2014 Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345**:** 1250091.

Pophaly, S. D., and A. Tellier, 2015 Population level purifying selection and gene expression shape subgenome evolution in maize. Molecular biology and evolution 32**:** 3226-3235.

Renny-Byfield, S., J. P. Gallagher, C. E. Grover, E. Szadkowski, J. T. Page *et al.*, 2014 Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. Genome biology and evolution 6**:** 559-571.

Rerie, W. G., K. A. Feldmann and M. D. Marks, 1994 The GLABRA2 gene encodes a homeo domain protein required for normal trichome development in Arabidopsis. Genes & development 8**:** 1388-1399.

Roche, 2001 Roche Applied Science Technical Note No. LC 13/2001, pp.

Rodriguez, M. A., D. Vermaak, J. J. Bayes and H. S. Malik, 2007 Species-specific positive selection of the male-specific lethal complex that participates in dosage compensation in Drosophila. Proceedings of the National Academy of Sciences 104**:** 15412-15417.

Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light *et al.*, 2007 Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 308**:** 58-73.

Samuel Yang, S., F. Cheung, J. J. Lee, M. Ha, N. E. Wei *et al.*, 2006 Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. The Plant Journal 47**:** 761-775.

Sato, Y., and H. Kudoh, 2015 Tests of associational defence provided by hairy plants for glabrous plants of Arabidopsis halleri subsp. gemmifera against insect herbivores. Ecological Entomology 40**:** 269-279.

Sattler, M. C., C. R. Carvalho and W. R. Clarindo, 2016 The polyploidy and its key role in plant breeding. Planta 243**:** 281-296.

Schellmann, S., A. Schnittger, V. Kirik, T. Wada, K. Okada *et al.*, 2002 TRIPTYCHON and CAPRICE mediate lateral inhibition during trichome and root hair patterning in Arabidopsis. Embo j 21**:** 5036-5046.

Schnable, J. C., and M. Freeling, 2011 Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PloS one 6**:** e17855.

Schnable, J. C., N. M. Springer and M. Freeling, 2011 Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences 108**:** 4069-4074.

Segraves, K. A., 2017 The effects of genome duplications in a community context. New Phytologist 215**:** 57-69.

Sehrish, T., V. V. Symonds, D. E. Soltis, P. S. Soltis and J. A. Tate, 2015 Cytonuclear coordination is not immediate upon allopolyploid formation in Tragopogon miscellus (Asteraceae) allopolyploids. PLoS One 10**:** e0144339.

Sémon, M., and K. H. Wolfe, 2007 Consequences of genome duplication. Current opinion in genetics & development 17**:** 505-512.

Serna, L., and C. Martin, 2006 Trichomes: different regulatory networks lead to convergent structures. Trends in plant science 11**:** 274-280.

Siena, L. A., J. P. A. Ortiz, O. Calderini, F. Paolocci, M. E. Cáceres *et al.*, 2016 An apomixis-linked ORC3-like pseudogene is associated with silencing of its functional homolog in apomictic Paspalum simplex. Journal of Experimental Botany 67**:** 1965-1978.

Suo, J., X. Liang, L. Pu, Y. Zhang and Y. Xue, 2003 Identification of GhMYB109 encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (Gossypium hirsutum L.). Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1630**:** 25-34.

Szymanski, D. B., R. A. Jilk, S. M. Pollock and M. D. Marks, 1998 Control of GL2 expression in Arabidopsis leaves and trichomes. Development 125**:** 1161-1171.

Thomas, B. C., B. Pedersen and M. Freeling, 2006 Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res 16**:** 934-946.

Thornton, B., and C. Basu, 2011 Real-time PCR (qPCR) primer design using free online software. Biochem Mol Biol Educ 39**:** 145-154.

Throude, M., S. Bolot, M. Bosio, C. Pont, X. Sarda *et al.*, 2009 Structure and expression analysis of rice paleo duplications. Nucleic acids research 37**:** 1248-1259.

Tsuji, J., and L. Coe, 2013 The glabra1 mutation affects the stomatal patterning of Arabidopsis thaliana rosette leaves. bios 84**:** 92-97.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth *et al.*, 2012 Primer3--new capabilities and interfaces. Nucleic acids research 40**:** e115-e115.

Van de Peer, Y., E. Mizrachi and K. Marchal, 2017 The evolutionary significance of polyploidy. Nature Reviews Genetics 18**:** 411-424.

VanderSluis, B., J. Bellay, G. Musso, M. Costanzo, B. Papp *et al.*, 2010 Genetic interactions reveal the evolutionary trajectories of duplicate genes. Molecular systems biology 6**:** 429.

Vanin, E. F., 1985 Processed pseudogenes: characteristics and evolution. Annual review of genetics 19**:** 253-272.

Videvall, E., N. Sletvold, J. Hagenblad, J. Ågren and B. Hansson, 2015 Strong Maternal Effects on Gene Expression in Arabidopsis lyrata Hybrids. Molecular Biology and Evolution 33**:** 984-994.

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

Walker, A. R., P. A. Davison, A. C. Bolognesi-Winfield, C. M. James, N. Srinivasan *et al.*, 1999 The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. The Plant cell 11**:** 1337-1350.

Wang, J., L. Tian, H.-S. Lee, N. E. Wei, H. Jiang *et al.*, 2006 Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics 172**:** 507-517.

Wang, S., and K. L. Adams, 2015 Duplicate gene divergence by changes in microRNA binding sites in Arabidopsis and Brassica. Genome biology and evolution 7**:** 646-655.

Wang, S., J.-W. Wang, N. Yu, C.-H. Li, B. Luo *et al.*, 2004 Control of plant trichome development by a cotton fiber MYB gene. The Plant Cell 16**:** 2323-2334.

Wang, X., Q. Dong, X. Li, A. Yuliang, Y. Yu *et al.*, 2017 Cytonuclear variation of Rubisco in synthesized rice hybrids and allotetraploids. The Plant Genome 10**:** plantgenome2017.2005.0041.

Wang, Z., J. Yang, F. Cheng, P. Li, X. Xin *et al.*, 2022 Subgenome dominance and its evolutionary implications in crop domestication and breeding. Horticulture Research 9**:** uhac090.

Wang, Z., Z. Yang and F. Li, 2019 Updates on molecular mechanisms in the development of branched trichome in Arabidopsis and nonbranched in cotton. Plant Biotechnology Journal 17**:** 1706-1722.

Wei, Z., Y. Cheng, C. Zhou, D. Li, X. Gao *et al.*, 2019 Genome-wide identification of direct targets of the TTG1–bHLH–MYB complex in regulating trichome formation and flavonoid accumulation in Arabidopsis thaliana. International Journal of Molecular Sciences 20**:** 5014.

Wu, X., and X. Qi, 2010 Genes encoding hub and bottleneck enzymes of the Arabidopsismetabolic network preferentially retain homeologs through whole genome duplication. BMC Evolutionary Biology 10**:** 145.

Wu, Y., Y. Sun, X. Wang, X. Lin, S. Sun *et al.*, 2016 Transcriptome shock in an interspecific F1 triploid hybrid of Oryza revealed by RNA sequencing. Journal of Integrative Plant Biology 58**:** 150-164.

Yan, M., X. Liu, C. Guan, L. Liu, J. Xiang *et al.*, 2014 Cloning of TTG1 gene and PCR identification of genomes A, B and C in Brassica species. Genetica 142**:** 169-176.

Yang, J., D. Liu, X. Wang, C. Ji, F. Cheng *et al.*, 2016 The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. Nature genetics 48**:** 1225-1232.

Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen *et al.*, 2012 Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13**:** 134.

Yim, W. C., B.-M. Lee and C. S. Jang, 2009 Expression diversity and evolutionary dynamics of rice duplicate genes. Molecular Genetics and Genomics 281**:** 495-495.

Yogeeswaran, K., C. Voelckel, S. Joly and P. B. Heenan, 2011 Pachycladon. 227-249.

Yoo, M. J., E. Szadkowski and J. F. Wendel, 2013 Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110**:** 171-180.

Zhang, B., D. Chopra, A. Schrader and M. Hülskamp, 2019 Evolutionary comparison of competitive protein-complex formation of MYB, bHLH, and WDR proteins in plants. Journal of Experimental Botany 70**:** 3197-3209.

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

Zhang, L., J. He, H. He, J. Wu and M. Li, 2021 Genome-wide unbalanced expression bias and expression level dominance toward Brassica oleracea in artificially synthesized intergeneric hybrids of Raphanobrassica. Horticulture Research 8.

Zhang, M., X.-K. Liu, W. Fan, D.-F. Yan, N.-S. Zhong *et al.*, 2018 Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from Camellia. Genome 61**:** 477-485.

Zhao, B., L. Liu, D. Tan and J. Wang, 2010 Analysis of phylogenetic relationships of Brassicaceae species based on Chs sequences. Biochemical Systematics and Ecology 38**:** 731-739.

Zhao, M., K. Morohashi, G. Hatlestad, E. Grotewold and A. Lloyd, 2008 The TTG1-bHLH-MYB complex controls trichome cell fate and patterning through direct targeting of regulatory loci. Development 135**:** 1991-1999.

Zheng, D., W. Ye, Q. Song, F. Han, T. Zhang *et al.*, 2016a Histone modifications define expression bias of homoeologous genomes in allotetraploid cotton. Plant physiology 172**:** 1760-1771.

Zheng, K., H. Tian, Q. Hu, H. Guo, L. Yang *et al.*, 2016b Ectopic expression of R3 MYB transcription factor gene OsTCL1 in Arabidopsis, but not rice, affects trichome and root hair formation. Scientific reports 6**:** 19254.

Zhou, L., K. Zheng, X. Wang, H. Tian, X. Wang *et al.*, 2014 Control of trichome formation in Arabidopsis by poplar single-repeat R3 MYB transcription factors. Frontiers in Plant Science 5**:** 262.

A4. APPENDIX

Following are the standard curves obtained for all the primer combinations used in the study. The results were generated in the LightCycler 480 software 1.5.1.62 SP3. Primer efficiencies were calculated using these standard curves.

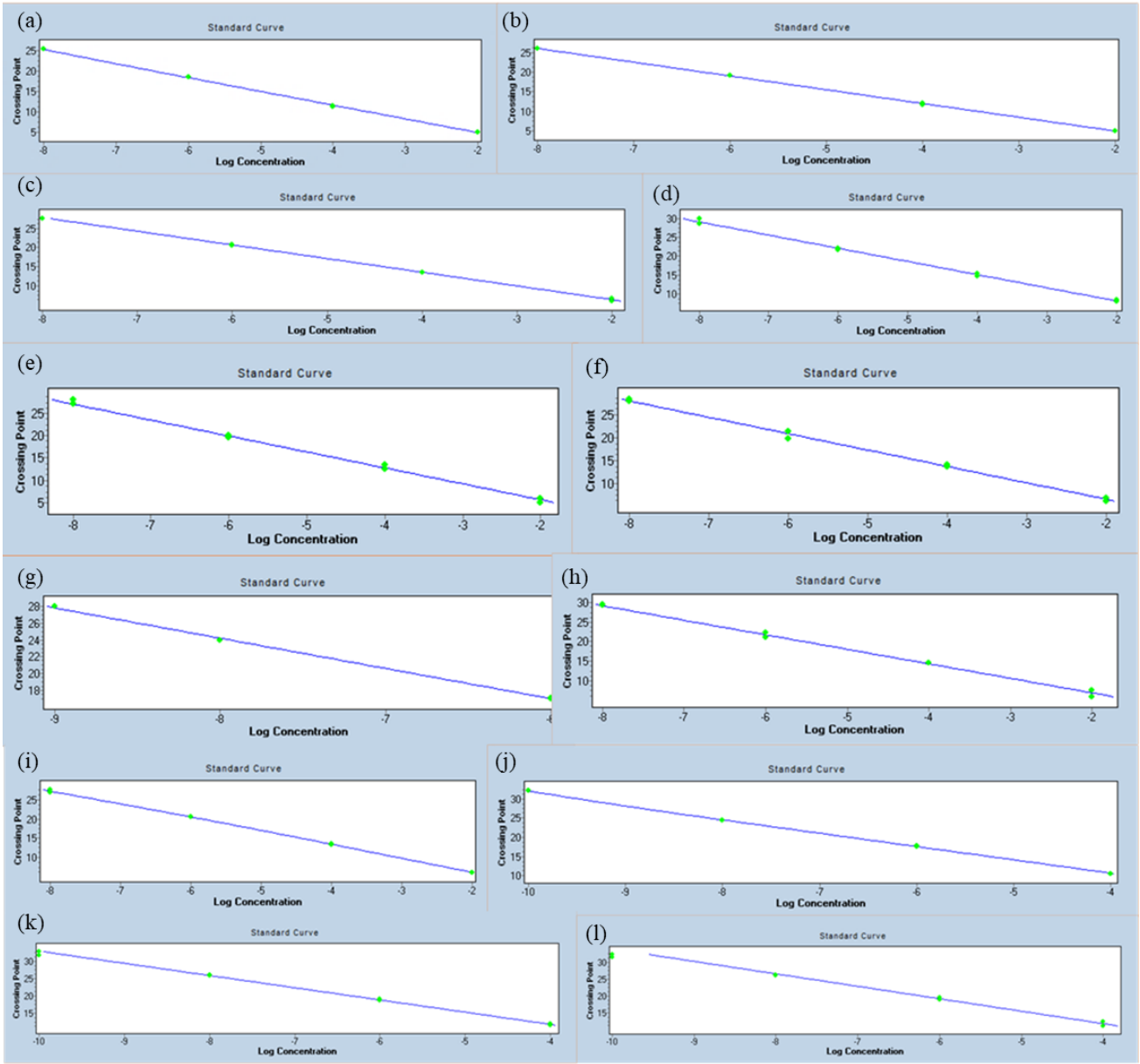
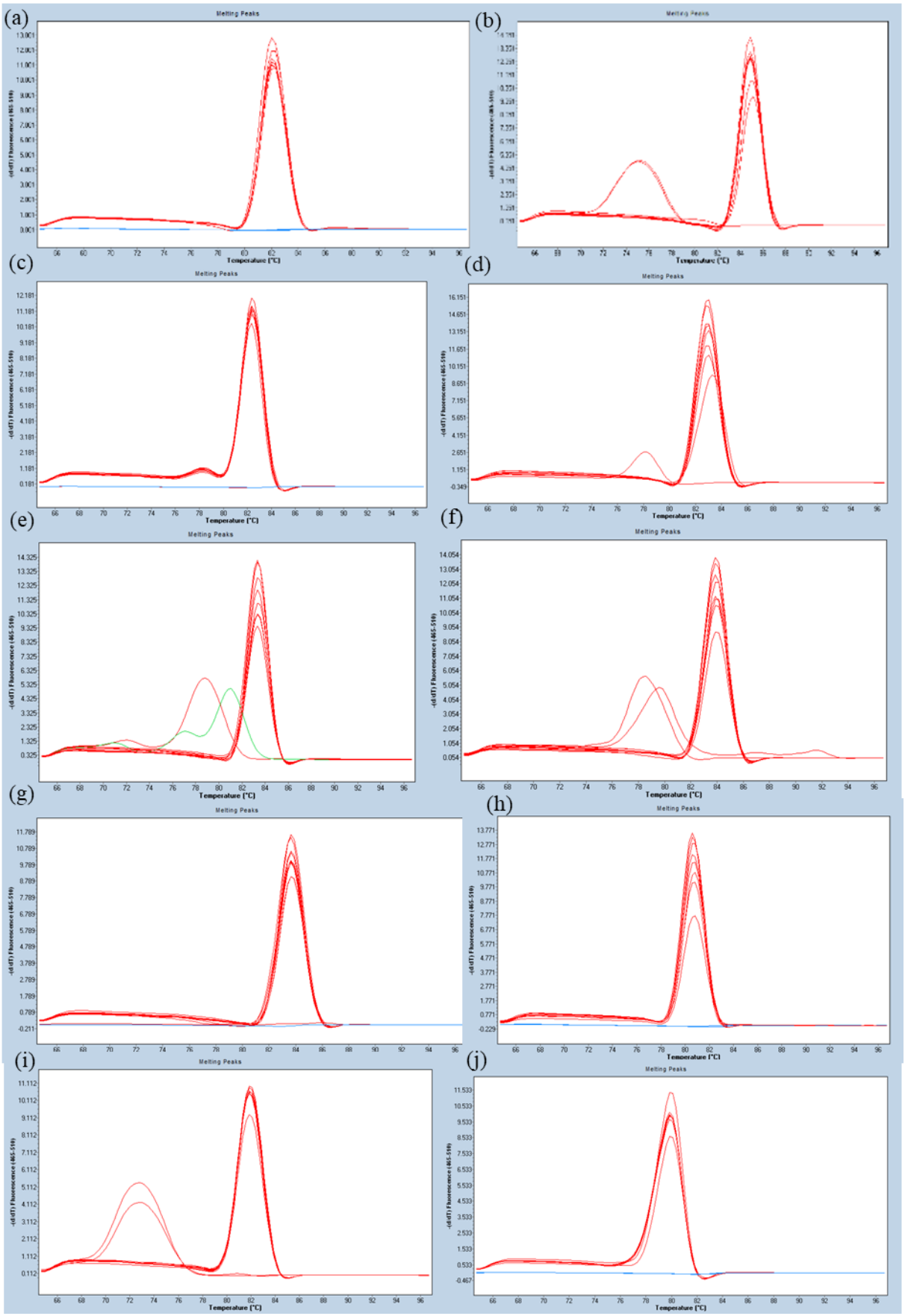


Figure A4.1: Standard curves for gene primer efficiencies (a) *TTG1*-Type1-1920F×1922R, (b) *TTG1*-Type2-1923F×1925R, (c) *GL1*-Type1-1994F×2004R, (d) *GL1*-Type2-1994F×2004R, (e) *GL3*-Type1-2235F×2236R, *P.enysii*, *P.fastigiatum* and *P.cheesemanii*- (f) *GL3*-Type1-*P.novae-zealandiae*- 2235F×2237R, (g) *GL3*-Type2- 2292F×2293R, (h) *EGL3*-Type1-2252F×2253R, (i) *EGL3*-Type2- 2254F×2255R, (j) Ref1-*CAC*- 1928F×1930R, (k) Ref2-*TIP41*-1929F×1931R and (l) Ref3-*PPR*-2003F×2012R. Cp (Crossing point) values in Y-axis are plotted against log concentrations of dilution series of PCR products.

Primer specificities were evaluated by the melt curves given below.


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
Melting Peaks
Temperature (°C)
-(d/dT) Fluorescence (465-510)

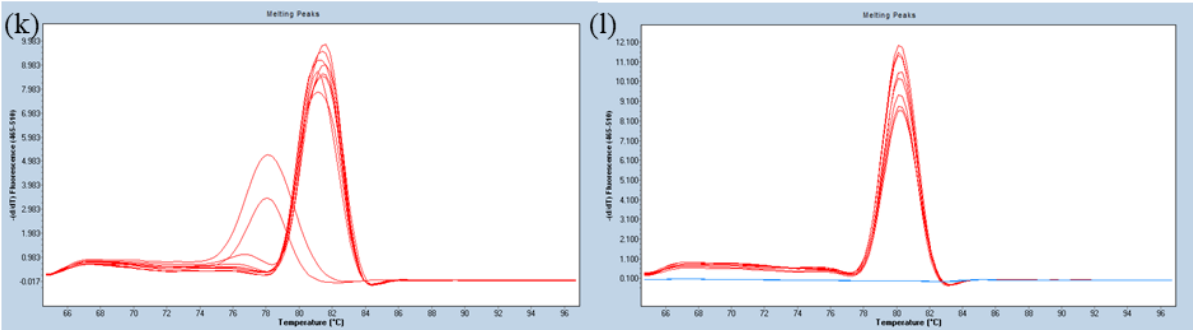


Figure A4.2: Melt curves of primer products; (a) *TTG1-Type1*-1920F×1922R, (b) *TTG1-Type2*-1923F×1925R, (c) *GL1-Type1*-1994F×2004R, (d) *GL1-Type2*-1994F×2004R, (e) *GL3-Type1-P.en*, *P.fs* and *P.ch*- 2235F×2236R, (f) *GL3-Type1-P.nz*- 2235F×2237R, (g) *GL3-Type2*-2292F×2293R, (h) *EGL3-Type1*- 2252F×2253R, (i) *EGL3-Type2*- 2254F×2255R, (j) Ref1-*CAC*- 1928F×1930R, (k) Ref2-*TIP41*-1929F×1931R and (l) Ref3-*PPR*-2003F×2012R. Non-specific amplifications in (b) around 75 °C, (d) 78.19 °C, (e) 77-81 °C, (f) around 79 °C, (i) around 73 °C and (k) around 78 °C are from negative water controls (NTC) and distinguishable from unique product specific melt curves, thus not effecting the primer specificities.

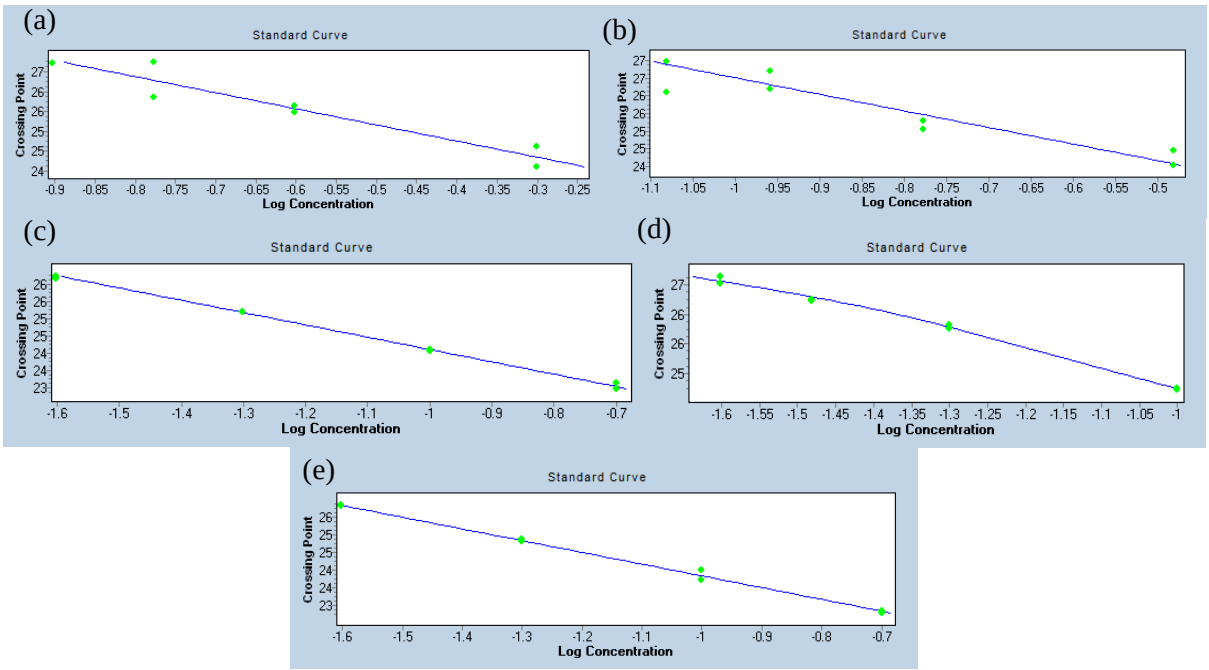


Figure A4.3: Standard curves for cDNA dilution series of (a) *P.enysii* seedling, (b) *P.fastigiatum* mature leaf, (c) *P.novae-zealandiae* mature leaf, (d) *P.cheesemanii* seedling, and (e) *P.cheesemanii* mature leaf.

Table A4.1: *Pachycladon* cDNA dilution series used with TTG1-T1-1920F×1922R primer pair and their primer efficiencies. RT controls used (No RT templates) are not shown here.

cDNA template	RNA concentration (ng/µL)	Dilution factor (fold)	Final concentration (ng/µL)	Cp	Primer efficiency
Y-P.en-cDNA-1	168	2	10	24.36	1.758
		4	5	25.56	
		6	3.33	26.31	
		8	2.5	26.73	
M-P.fs-cDNA-2	265.6	3	10.54	24.25	1.629
		6	5.27	25.17	
		9	3.51	26.45	
		12	2.64	26.54	
M-P.nz-cDNA-1	774.4	5	18.44	23.05	1.902
		10	9.23	24.07	
		20	4.61	25.2	
		40	2.31	26.19	
M-P.nz-cDNA-2	1000	5	23.8	22.81	1.996
		10	11.9	23.84	
		20	5.95	24.85	
		40	2.98	25.82	
Y-P.ch-cDNA-1	1000	10	11.9	24.73	1.922
		20	5.95	25.79	
		30	3.97	26.24	
		40	2.98	26.59	

Table A4.2: *Pachycladon* expression data (Cp values) for the genes of interest and reference genes. '#No RT' column indicates whether the data is from a negative template made without using reverse transcriptase ('0') or from a target sample started from a template amplified using reverse transcriptase ('1'). '#IPC' (InterPlate Control) indicates whether the sample has been used as an interplate control among different runs in the experiment. '#PCR repeat' shows the biological replicates. '#Plate' indicates the plate/experiment number the sample was run. The table was prepared (keeping empty cells in between) in an input file format that GenEx 7.0 could read.

Sample	GOI_TTG1T1	GOI_TTG1T2	GOI_GL1T1	GOI_GL1T2	GOI_GL3T1	GOI_GL3T2	GOI_EGL3T1	GOI_EGL3T2	Ref1_CAC	Ref2_TIP41	Ref3_PPR	#No RT	#IPC	#PCR repeat	#Plate
Y-P.en cDNA1	24.54								23.05	23.66	27.29	0	0	1	1
Y-P.en cDNA1	25.23								23.09	23.72	27.52	0	0	1	1
Y-P.en cDNA1	24.98											0	0	1	1
Y-P.en cDNA2	25.03								23.03	23.96	28.17	0	0	2	1
Y-P.en cDNA2	25.28								23.03	23.98	28.26	0	0	2	1
Y-P.en cDNA2	25.07											0	0	2	1
Y-P.en cDNA3	24.73								22.16	23.47	27.33	0	0	3	1
Y-P.en cDNA3	24.77								22.99	23.54	27.56	0	0	3	1
Y-P.en cDNA3	24.71											0	0	3	1
Y-P.en cDNA1		24.82										0	0	4	3
Y-P.en cDNA1		24.49										0	0	4	3
Y-P.en cDNA1		25.28										0	0	4	3
Y-P.en cDNA2		25.44										0	0	5	3
Y-P.en cDNA2		25.09										0	0	5	3
Y-P.en cDNA2		25.02										0	0	5	3
Y-P.en cDNA3		24.60										0	0	6	3

Y-P.en cDNA3		24.42										0	0	6	3
Y-P.en cDNA3		24.43										0	0	6	3
Y-P.en cDNA1			31.29									0	0	7	4
Y-P.en cDNA1			31.51									0	0	7	4
Y-P.en cDNA1			31.60									0	0	7	4
Y-P.en cDNA2			29.56									0	0	8	4
Y-P.en cDNA2			29.88									0	0	8	4
Y-P.en cDNA2			30.31									0	0	8	4
Y-P.en cDNA3			29.77									0	0	9	4
Y-P.en cDNA3			29.59									0	0	9	4
Y-P.en cDNA3			29.91									0	0	9	4
Y-P.en cDNA1				31.64	27.85	29.61						0	0	10	5
Y-P.en cDNA1				31.9	27.57	29.88						0	0	10	5
Y-P.en cDNA1				31.98	27.72	29.79						0	0	10	5
Y-P.en cDNA2				30.99	27.45	30.21						0	0	11	5
Y-P.en cDNA2				30.87	27.49	30.57						0	0	11	5
Y-P.en cDNA2				31.89	27.44	29.78						0	0	11	5
Y-P.en cDNA3				31.6	26.74	29.46						0	0	12	5
Y-P.en cDNA3				30.64	26.72	29.06						0	0	12	5
Y-P.en cDNA3				30.99	26.74	29.24						0	0	12	5
Y-P.en cDNA1							29.98	28.43				0	0	13	6
Y-P.en cDNA1							29.3	28.2				0	0	13	6
Y-P.en cDNA1							29.56	28.53				0	0	13	6
Y-P.en cDNA2							28.58	26.85				0	0	14	6
Y-P.en cDNA2							28.68	27.04				0	0	14	6
Y-P.en cDNA2							28.78	27.01				0	0	14	6
Y-P.en cDNA3							28.69	27.28				0	0	15	6
Y-P.en cDNA3							28.32	27.1				0	0	15	6
Y-P.en cDNA3							28.58	27.13				0	0	15	6
P.fs cDNA1	24.85	25.79							23.32		27.72	0	0	16	2
P.fs cDNA1	24.91	26.50							23.44	23.77	27.81	0	0	16	2
P.fs cDNA1	24.95	26.03										0	0	16	2
P.fs cDNA2	24.80	25.55							23.12	23.20	27.19	0	0	17	2
P.fs cDNA2	24.88	24.94							23.16	23.21	26.97	0	0	17	2
P.fs cDNA2	24.87	25.95										0	0	17	2
P.fs cDNA3	25.06	25.25							23.16	23.89	27.63	0	0	18	2
P.fs cDNA3	25.02	26.26							23.09	23.97	27.73	0	0	18	2
P.fs cDNA3	25.11	26.00										0	0	18	2
P.fs cDNA1			33.29	33.55								0	0	19	4
P.fs cDNA1			34.47	33.91								0	0	19	4
P.fs cDNA1			32.87	32.90								0	0	19	4
P.fs cDNA2			30.73	35.66								0	0	20	4
P.fs cDNA2			30.95	34.48								0	0	20	4
P.fs cDNA2			30.94	40								0	0	20	4
P.fs cDNA3			32.27	32.86								0	0	21	4
P.fs cDNA3			32.59	34.51								0	0	21	4
P.fs cDNA3			32.19	34.55								0	0	21	4
P.fs cDNA1					26.66	30.02						0	0	22	5
P.fs cDNA1					26.62	30.16						0	0	22	5
P.fs cDNA1					26.51	29.98						0	0	22	5
P.fs cDNA2					25.24	29.91						0	0	23	5
P.fs cDNA2					25.72	30.23						0	0	23	5
P.fs cDNA2					24.91	30.35						0	0	23	5
P.fs cDNA3					25.74	30.01						0	0	24	5
P.fs cDNA3					25.91	30.03						0	0	24	5
P.fs cDNA3					25.81	29.8						0	0	24	5
P.fs cDNA1							30.56	29.51				0	0	25	6
P.fs cDNA1							30.32	30.19				0	0	25	6
P.fs cDNA1							30.57	30.18				0	0	25	6
P.fs cDNA2							28.91	28.03				0	0	26	6
P.fs cDNA2							29.05	28.14				0	0	26	6
P.fs cDNA2							28.81	27.9				0	0	26	6
P.fs cDNA3							28.34	28.46				0	0	27	6
P.fs cDNA3							28.32	28.29				0	0	27	6
P.fs cDNA3							28.14	28.15				0	0	27	6
P.nz cDNA1	24.61								23.46	23.82	29.30	0	0	28	2
P.nz cDNA1	24.60								23.51	23.80	29.30	0	0	28	2
P.nz cDNA1	24.64											0	0	28	2
P.nz cDNA2	24.98								23.69	24.16	29.83	0	0	29	2
P.nz cDNA2	25.02								23.83	24.44	29.88	0	0	29	2
P.nz cDNA2	24.95											0	0	29	2
P.nz cDNA3	24.77								23.25	23.75	29.08	0	0	30	2
P.nz cDNA3	24.72								23.45	23.98	29.53	0	0	30	2
P.nz cDNA3	24.72											0	0	30	2
P.nz cDNA1		24.66										0	0	31	3
P.nz cDNA1		24.61										0	0	31	3
P.nz cDNA1		24.47										0	0	31	3
P.nz cDNA2		25.23										0	0	32	3
P.nz cDNA2		25.06										0	0	32	3
P.nz cDNA2		25.11										0	0	32	3
P.nz cDNA3		24.82										0	0	33	3

P.nz cDNA3		24.87										0	0	33	3
P.nz cDNA3		24.73										0	0	33	3
P.nz cDNA1			32.25	34.22								0	0	34	4
P.nz cDNA1			33.12	35.55								0	0	34	4
P.nz cDNA1			32.42	33.29								0	0	34	4
P.nz cDNA2			34.30	32.97								0	0	35	4
P.nz cDNA2			33.09	34.60								0	0	35	4
P.nz cDNA2			33.50	33.99								0	0	35	4
P.nz cDNA3			33.90	34.09								0	0	36	4
P.nz cDNA3			33.70	35.65								0	0	36	4
P.nz cDNA3			40	33.87								0	0	36	4
P.nz cDNA1					25.43	29.66	29.54	29.89				0	0	37	6
P.nz cDNA1					26.17	29.89	29.78	29.76				0	0	37	6
P.nz cDNA1					25.77	29.78	29.81	29.85				0	0	37	6
P.nz cDNA2					26.47	29.65	29.64	31.08				0	0	38	6
P.nz cDNA2					26.42	29.32	29.53	32.56				0	0	38	6
P.nz cDNA2					26.45	29.71	29.74	31.71				0	0	38	6
P.nz cDNA3					26.04	29.43	29.11	31.23				0	0	39	6
P.nz cDNA3					25.97	29.61	29.05	31.85				0	0	39	6
P.nz cDNA3					26.03	29.48	29.19	31.8				0	0	39	6
Y-P.ch cDNA1	25.5								23.64	24.02	28.86	0	0	40	1
Y-P.ch cDNA1	25.63								23.67	24.02	28.82	0	0	40	1
Y-P.ch cDNA1	25.85											0	0	40	1
Y-P.ch cDNA2	25.34								23.42	23.95	28.7	0	0	41	1
Y-P.ch cDNA2	25.44								23.3	23.94	28.95	0	0	41	1
Y-P.ch cDNA2	25.46											0	0	41	1
Y-P.ch cDNA3	25.69								23.53	24.08	28.66	0	0	42	1
Y-P.ch cDNA3	25.54								23.61	24.12	28.8	0	0	42	1
Y-P.ch cDNA3	25.72											0	0	42	1
Y-P.ch cDNA1		25.30										0	0	43	3
Y-P.ch cDNA1		25.33										0	0	43	3
Y-P.ch cDNA1		25.66										0	0	43	3
Y-P.ch cDNA2		25.31										0	0	44	3
Y-P.ch cDNA2		25.53										0	0	44	3
Y-P.ch cDNA2		25.53										0	0	44	3
Y-P.ch cDNA3		25.69										0	0	45	3
Y-P.ch cDNA3		25.62										0	0	45	3
Y-P.ch cDNA3		25.64										0	0	45	3
Y-P.ch cDNA1			28.49	30.19								0	0	46	4
Y-P.ch cDNA1			28.50	29.93								0	0	46	4
Y-P.ch cDNA1			28.47	30.81								0	0	46	4
Y-P.ch cDNA2			28.66	30.76								0	0	47	4
Y-P.ch cDNA2			28.53	31.30								0	0	47	4
Y-P.ch cDNA2			28.24	30.43								0	0	47	4
Y-P.ch cDNA3			29.05	31.83								0	0	48	4
Y-P.ch cDNA3			28.83	31.52								0	0	48	4
Y-P.ch cDNA3			28.77	31.47								0	0	48	4
Y-P.ch cDNA1					26.92	34.46						0	0	49	5
Y-P.ch cDNA1					26.78							0	0	49	5
Y-P.ch cDNA1					26.47	32.72						0	0	49	5
Y-P.ch cDNA2					26.79	32.98						0	0	50	5
Y-P.ch cDNA2					26.55	33.32						0	0	50	5
Y-P.ch cDNA2					26.51	33.16						0	0	50	5
Y-P.ch cDNA3					26.91	34.69						0	0	51	5
Y-P.ch cDNA3					26.86	33.74						0	0	51	5
Y-P.ch cDNA3					26.93	34.89						0	0	51	5
Y-P.ch cDNA1							26.72	26.94				0	0	52	7
Y-P.ch cDNA1							26.95	26.97				0	0	52	7
Y-P.ch cDNA1							27.11	27.02				0	0	52	7
Y-P.ch cDNA2							26.51	26.47				0	0	53	7
Y-P.ch cDNA2							26.60	26.62				0	0	53	7
Y-P.ch cDNA2							26.75	26.51				0	0	53	7
Y-P.ch cDNA3							26.67	26.91				0	0	54	7
Y-P.ch cDNA3							26.69	26.98				0	0	54	7
Y-P.ch cDNA3							26.78	26.85				0	0	54	7
M-P.ch cDNA1	24.47	24.88							22.63	22.81	27.79	0	0	55	3
M-P.ch cDNA1	24.98	24.85							22.89	22.99	28.74	0	0	55	3
M-P.ch cDNA1	24.75	24.79										0	0	55	3
M-Pch cDNA2	24.89	24.81							22.90	22.94	28.16	0	0	56	3
M-Pch cDNA2	24.99	24.89							22.80	23.07	28.34	0	0	56	3
M-Pch cDNA2	24.99	24.93										0	0	56	3
M-P.ch cDNA3	25.67	25.56							23.22	23.47	28.16	0	0	57	3
M-P.ch cDNA3	25.65	24.92							23.24	23.55	28.27	0	0	57	3
M-P.ch cDNA3	25.14	25.49										0	0	57	3
M-P.ch cDNA1			36.08	33.28								0	0	58	4
M-P.ch cDNA1			35.97	33.91								0	0	58	4
M-P.ch cDNA1			40	34.44								0	0	58	4
M-Pch cDNA2			33.15	32.66								0	0	59	4
M-Pch cDNA2			34.95	33.64								0	0	59	4
M-Pch cDNA2			34.71	32.47								0	0	59	4
M-P.ch cDNA3			32.58	32.24								0	0	60	4

M-P.ch cDNA3			32.76	32.11								0	0	60	4
M-P.ch cDNA3			32.85	33.46								0	0	60	4
M-P.ch cDNA1					26.62	36.39						0	0	61	5
M-P.ch cDNA1					26.32	33.66						0	0	61	5
M-P.ch cDNA1					26.53	34.76						0	0	61	5
M-Pch cDNA2					25.49	33.65						0	0	62	5
M-Pch cDNA2					25.75	34.32						0	0	62	5
M-Pch cDNA2					25.79	36.01						0	0	62	5
M-P.ch cDNA3					25.9	33.6						0	0	63	5
M-P.ch cDNA3					25.93	32.55						0	0	63	5
M-P.ch cDNA3					25.95	33.13						0	0	63	5
M-P.ch cDNA1							26.32	28.27				0	0	64	7
M-P.ch cDNA1							26.34	28.10				0	0	64	7
M-P.ch cDNA1							26.49	28.35				0	0	64	7
M-Pch cDNA2							26.50	27.89				0	0	65	7
M-Pch cDNA2							26.51	28.18				0	0	65	7
M-Pch cDNA2							26.49	28.10				0	0	65	7
M-P.ch cDNA3							26.05	27.14				0	0	66	7
M-P.ch cDNA3							26.01	27.21				0	0	66	7
M-P.ch cDNA3							26.00	27.23				0	0	66	7
NTC_plate1										34.14		0	0	67	1
NTC_plate1										33.69		0	0	67	1
NTC_plate2		31.50									40	0	0	68	2
NTC_plate2		29.64									40	0	0	68	2
NTC_plate3		27.29							31.85	30.99		0	0	69	3
NTC_plate3		30.79								35.80		0	0	69	3
NTC_plate4						37.59						0	0	70	4
NTC_plate4						40.00				36.76		0	0	70	4
NTC_plate5				40	35.2	34.73				36.41		0	0	71	5
NTC_plate5				40	34.53	34.5				34.21		0	0	71	5
NTC_plate6					37.21	35.67		34.57		40		0	0	72	6
NTC_plate6					36.34	34.33		33.63				0	0	72	6
NTC_plate7								32.67		40.00		0	0	73	7
NTC_plate7								32.07		40.00		0	0	73	7
Y-P.en cDNA1_no RT												1	0	74	1
Y-P.en cDNA1_no RT												1	0	74	1
Y-P.en cDNA2_no RT												1	0	75	1
Y-P.en cDNA2_no RT												1	0	75	1
Y-P.en cDNA3_no RT	34.84											1	0	76	1
Y-P.en cDNA3_no RT												1	0	76	1
P.fs cDNA1_no RT	33.32											1	0	77	2
P.fs cDNA1_no RT	33.48											1	0	77	2
P.fs cDNA2_no RT												1	0	78	2
P.fs cDNA2_no RT	36.28											1	0	78	2
P.fs cDNA3_no RT												1	0	79	2
P.fs cDNA3_no RT	35.49											1	0	79	2
P.nz cDNA1_no RT	35.80											1	0	80	2
P.nz cDNA1_no RT	33.29											1	0	80	2
P.nz cDNA2_no RT	34.88											1	0	81	2
P.nz cDNA2_no RT	34.90											1	0	81	2
P.nz cDNA3_no RT												1	0	82	2
P.nz cDNA3_no RT	34.83											1	0	82	2
Y-P.ch cDNA1_no RT	33.85											1	0	83	1
Y-P.ch cDNA1_no RT	34.86											1	0	83	1
Y-P.ch cDNA2_no RT	34.86											1	0	84	1
Y-P.ch cDNA2_no RT	34.81											1	0	84	1
Y-P.ch cDNA3_no RT	34.61											1	0	85	1
Y-P.ch cDNA3_no RT	35.82											1	0	85	1

IPC_Y_P.en cDNA2_plate1										23.96		0	1	86	1
IPC_Y_P.en cDNA2_plate1										23.98		0	1	86	1
IPC_Y_P.ch cDNA2_plate1										23.95		0	1	87	1
IPC_Y_P.ch cDNA2_plate1										23.94		0	1	87	1
IPC_Y_P.en cDNA2_plate2										23.83		0	1	88	2
IPC_Y_P.en cDNA2_plate2										23.47		0	1	88	2
IPC_Y_P.ch cDNA2_plate2										24.00		0	1	89	2
IPC_Y_P.ch cDNA2_plate2										24.49		0	1	89	2
IPC_Y_P.en cDNA2_plate3										23.56		0	1	90	3
IPC_Y_P.en cDNA2_plate3										23.64		0	1	90	3
IPC_Y_P.ch cDNA2_plate3										23.60		0	1	91	3
IPC_Y_P.ch cDNA2_plate3										23.65		0	1	91	3
IPC_P.fs cDNA2_plate3										22.89		0	1	92	3
IPC_P.fs cDNA2_plate3										22.99		0	1	92	3
IPC_P.nz cDNA2_plate3										23.58		0	1	93	3
IPC_P.nz cDNA2_plate3										23.77		0	1	93	3
IPC_Y_P.en cDNA2_plate4										23.66		0	1	94	4
IPC_Y_P.en cDNA2_plate4										23.75		0	1	94	4
IPC_Y_P.ch cDNA2_plate4										23.44		0	1	95	4
IPC_Y_P.ch cDNA2_plate4										23.27		0	1	95	4
IPC_Y_P.en cDNA2_plate5										23.94		0	1	96	5
IPC_Y_P.en cDNA2_plate5										23.82		0	1	96	5
IPC_Y_P.en cDNA2_plate5										23.85		0	1	96	5
IPC_Y_P.ch cDNA1_plate5										23.75		0	1	97	5
IPC_Y_P.ch cDNA1_plate5										23.62		0	1	97	5
IPC_Y_P.ch cDNA1_plate5										23.72		0	1	97	5
IPC_Y_P.ch cDNA2_plate5										23.79		0	1	98	5
IPC_Y_P.en cDNA2_plate6										23.76		0	1	99	6
IPC_Y_P.en cDNA2_plate6										23.87		0	1	99	6
IPC_Y_P.en cDNA2_plate6										23.69		0	1	99	6
IPC_Y_P.ch cDNA2_plate6										23.76		0	1	100	6
IPC_Y_P.ch cDNA2_plate6										23.62		0	1	100	6
IPC_Y_P.ch cDNA2_plate6										23.7		0	1	100	6
IPC_Y_P.en cDNA2_plate7										23.88		0	1	101	7
IPC_Y_P.en cDNA2_plate7										24.10		0	1	101	7
IPC_Y_P.en cDNA2_plate7										23.58		0	1	101	7
IPC_Y_P.ch cDNA2_plate7										24.88		0	1	102	7
IPC_Y_P.ch cDNA2_plate7										24.47		0	1	102	7
IPC_Y_P.ch cDNA2_plate7										24.46		0	1	102	7
#Gene type	1	2	1	2	1	2	1	2	0	0	0	0	0	0	0

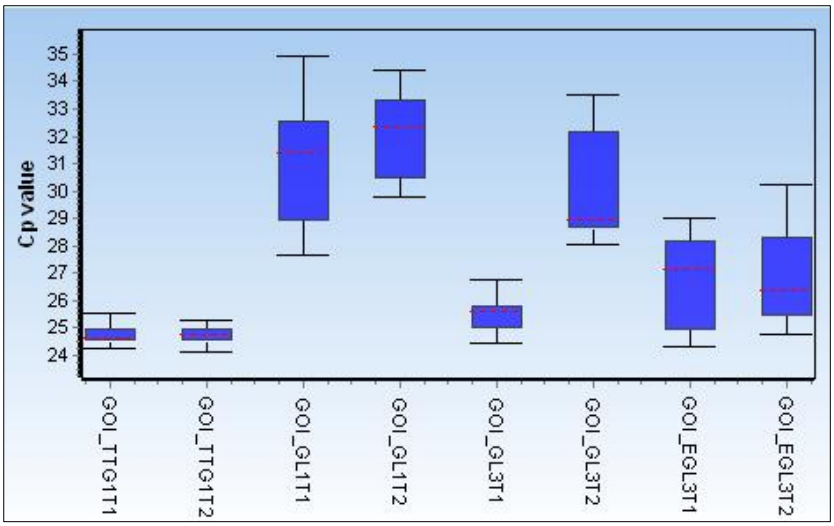


Figure A4.4: Distribution of Cp values of homeologous genes.

Table A4.3: Gene expression analysis using qRAT (Flatschacher et al 2022) to validate the results obtained. FC – Fold change, sd – standard deviation, *P. en- P. enysii*. *P. ch – P. cheesemanii*, *P. fs – P. fastigiatum* and *P. nz – P. novae-zealandiae*. The significant values are highlighted in brown that are less than the p value threshold 0.05.

Sample	**Difference of Type 1 compared to Type 2 homeolog (Type 1 - Type 2)**				
	Gene	t.test	p.value	adj.p.value	significance
P.en young seedling	EGL3	-3.9635	0.0019	0.0132	*
	GL1	2.2483	0.0441	0.3089	
	GL3	10.0119	<0.0001	<0.0001	****
	TTG1	0.0078	0.9939	>0.9999	
P. ch young seedling	EGL3	-0.1354	0.8946	>0.9999	
	GL1	9.6428	<0.0001	<0.0001	****
	GL3	27.3824	<0.0001	<0.0001	****
	TTG1	-0.8406	0.4174	>0.9999	
P.ch mature leaf	EGL3	11.6184	<0.0001	<0.0001	****
	GL1	-1.4	0.1962	>0.9999	
	GL3	21.8346	<0.0001	<0.0001	****
	TTG1	0.145	0.8872	>0.9999	
P. fs mature leaf	EGL3	-1.1564	0.269	>0.9999	
	GL1	3.1566	0.0108	0.0755	
	GL3	11.0508	<0.0001	<0.0001	****
	TTG1	3.2645	0.0064	0.0449	*
P. nz mature leaf	EGL3	-2.2712	0.0443	0.3098	
	GL1	-1.4497	0.1778	>0.9999	
	GL3	-16.6877	<0.0001	<0.0001	****
	TTG1	-0.2918	0.7759	>0.9999	

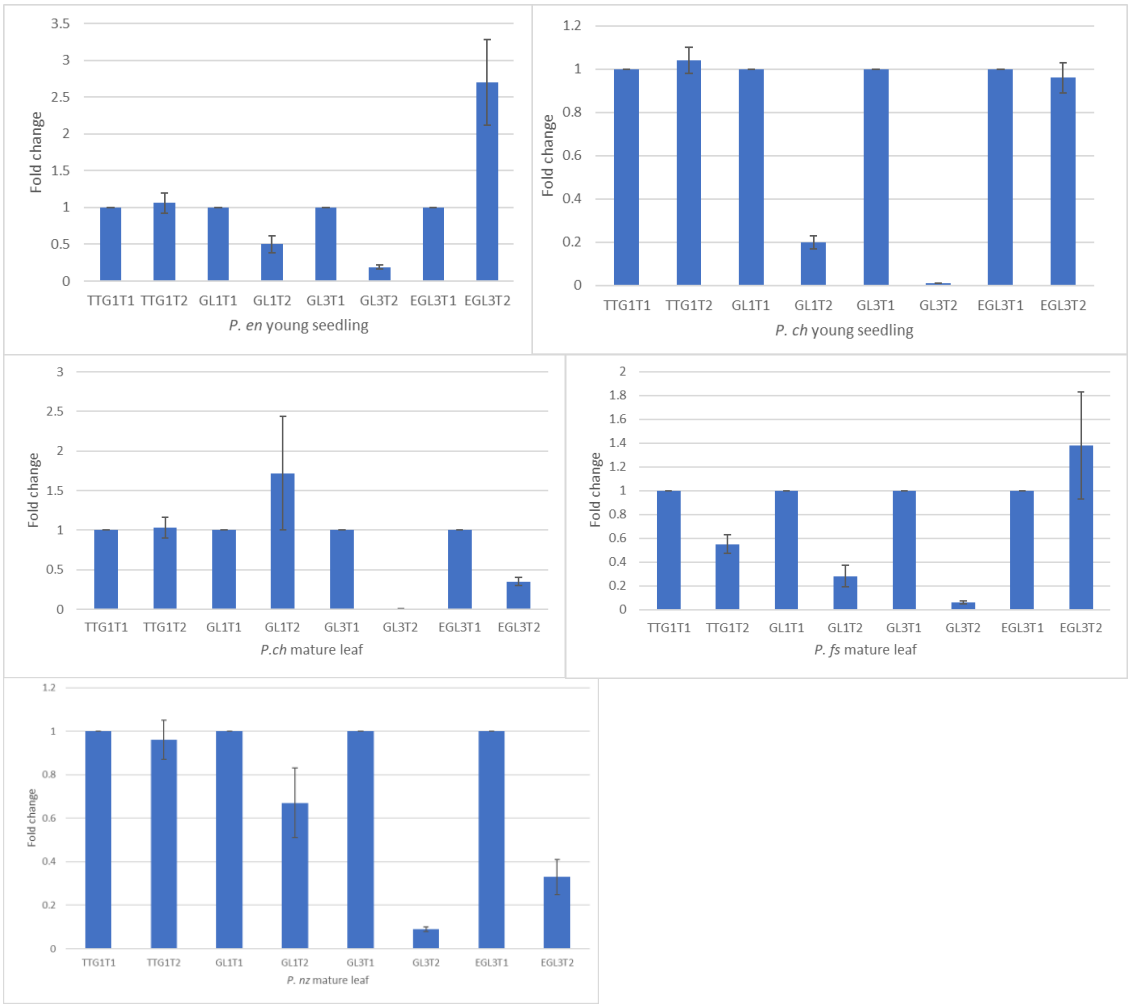


Figure A4.5: qRAT Gene expression results of *TTG1*, *GL1*, *GL3* and *EGL3* homeologs in *P. enysii* (*P. en*) seedling, *P. cheesemanii* (*P. ch*) seedling, *P. ch* mature leaf, *P. fastigiatum* (*P. fs*) mature leaf and *P. novae-zealandiae* (*P. nz*) mature leaf.

A4.1 Statement of contribution

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Thilini Madara Warusawithana
Name and title of main supervisor:	Dr. V. Vaughan Symonds
In which chapter is the manuscript/published work?	Chapter 4
What percentage of the manuscript/published work was contributed by the student?	85%

Describe the contribution that the student has made to the manuscript/published work:

Thilini worked with Dr. Vaughan Symonds to develop the research questions and objectives, and to design the experimental setup and data collection methods. Thilini conducted a comprehensive review of existing literature related to relative gene expression analysis. Thilini conducted analyses and interpreted the results incorporating the feedback from Dr. Symonds and Dr. Anja Schiemann. Thilini drafted the chapter and incorporated the feedback from Dr. Symonds and co-supervisor Dr. Matthieu Vignes.

Please select one of the following three options:

- ○ The manuscript/published work is published or in press
  Please provide the full reference of the research output:
- ○ The manuscript is currently under review for publication
  Please provide the name of the journal:
- ◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Student's signature:	Thilini Warusawithana Digitally signed by Thilini Warusawithana Date: 2024.10.25 15:13:08 +13'00'	Main supervisor's signature:	Vaughan Symonds Digitally signed by Vaughan Symonds Date: 2024.10.29 11:44:35 +13'00'

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

1

Doctoral Research Committee
July 2022

5.0 CHAPTER FIVE

Thesis summary and future directions

Whole-genome duplication (WGD) is a characteristic of almost all primary lineages of land plants (Clark & Donoghue 2018). Although duplicated genomes have been studied in the contexts of individual genes or genomes in the past from genetic, genomic, and epigenetic perspectives (Jackson & Chen 2010, Soltis et al 2010, Soltis & Soltis 2000), comparatively little work has been done at the level of genetic complexes or pathways. To this end, I studied the molecular evolution and expression of gene duplicates (homeologs) in the trichome initiation pathway, one of the conserved epidermal cell fate pathways well-annotated in Brassicaceae (Szymanski et al 2000, Walker & Marks 2000, Wang et al 2019, Xuan et al 2020). As the study system, we have chosen the New Zealand native allopolyploid genus, *Pachycladon*, owing to its characteristics such as having highly diverged progenitors, wide morphological and habitat diversities, and the close phylogenetic relationship of one of the subgenomes with *A. thaliana* (Dong et al 2019, Joly et al 2009, Yogeeswaran et al 2011). Thus, we have focused on the molecular evolution of the WD40-R2R3 MYB-bHLH complex in the trichome initiation pathway represented by the genes *TTG1*, *GL1*, *GL3*, and *EGL3* in *Pachycladon*, in our study, including four *Pachycladon* species, *P. enysii*, *P. fastigiatum*, *P. novae-zealandiae*, and *P. cheesemanii*, to represent diversity within the *Pachycladon* lineage.

The whole workflow of the study was dependent on the first objective to characterize the homeologs of the genes of interest (*TTG1*, *GL1*, *GL3*/*EGL3*) and to assign each copy to one hypothetical ancestor or the other to assemble the 'ancestral' pathways. After the homeolog sequences were isolated and grouped with other homeologs of the same type, they were analyzed for molecular evolution by examining the presence or absence of homeologs, and mutations in the sequences, and trying to identify mutations that potentially affect protein structure and function. Next, gene expression profiles were generated for the two homeologs of each gene to identify any expression differences between homeologs. The ultimate goal of

these analyses was to interpret the molecular evolution of the duplicated trichome initiation pathway by relating our results to previously addressed attributes important in assessing molecular evolution, such as pleiotropy, epistasis, functional overlap, gene dosage balance, subgenome dominance and selection pressures towards functional divergence. The results presented in this thesis are summarised and future directions are considered here.

5.1 Reconstruction of the two progenitor trichome development pathways

Although *Pachycladon* is known to be an allopolyploid, the progenitors are unknown. Therefore, the first step was attempting to sequence the complete coding regions of homeologs belonging to the four core genes of interest forming the trichome initiation complex in our four study species. In agreement with previous works (Joly et al 2009), a high degree of sequence variation between homeologs of each gene (9.3-11.4%) was identified here. This is in line with the idea that *Pachycladon* represents one of the most extreme examples of an allopolyploid lineage with highly diverged progenitors. Joly et al (2009) with an estimated divergence time of ~6.4 myr before the hybridization event.

Our phylogenetic gene trees were based on only a single locus and a limited number of taxa compared to the highly resolved *Brassicaceae* phylogenetic trees constructed recently (Hendriks et al 2023, Huang et al 2015, Nikolov et al 2019), which explains variation in the placement of some Brassicaceae species from one gene tree to another and to those in the much larger phylogenetic trees published by others. However, our focus was to identify the relative placements of the two homeologs of each gene in the context of the rest of the family, rather than deducing the phylogeny of all homologs used. We hypothesized that the Type 1 copy of each gene would cluster with orthologs from a particular Brassicaceae lineage, thus all Type 1 copies originate from one parent, and all Type 2 copies are derived from the other parent. Results showed that the placements of Type 1 and Type 2 homeologs were fairly consistent across the different gene trees produced. That is, the Type 1 homeolog of all four genes was grouped in Lineage I, with *Capella*, *Camelina*, and *Arabidopsis* homologs, and the Type 2

homeolog of all genes generally formed a sister group to the Lineage I species similar to work on other genes or markers in previous studies (Huang et al 2015, Joly et al 2009, Zhao et al 2010). So, the interpretation is that all Type 1 copies trace back to one ancestor of *Pachycladon* and all Type 2 copies trace back to the other ancestor supported our hypothesis. Importantly, this pieced together the 'ancestral' trichome initiation genetic pathways representing the diverged *TTG1-GL1-GL3/EGL3* complexes. It is important to point out that the gene sequences isolated from contemporary *Pachycladon* plants have, of course, evolved since the initial hybridization event, so do not precisely resemble the ancestral copies; however, polymorphisms that consistently differentiate Type 1 and Type 2 homeologs for a given gene likely represent polymorphisms that differentiated the progenitor homologs prior to hybridization.

5.2 Molecular evolution of the duplicated genetic pathway

After assembling the ancestral pathways, the molecular evolution of each homeolog was characterized using sequence variation, Ka/Ks analyses, and analyses of inferred protein sequences to gain insights into the duplicated genetic pathway. Our predictions and hypotheses suggested the Type 1 subgenome is biased, based on more potentially deleterious mutations observed in many Type 2 gene copies' sequences (except *TTG1* and *EGL3*) compared to Type 1 intact sequences identified in Chapter 1.

Generally, the nucleotide diversity of Type 1 homeologs was less than the Type 2 homeologs indicating more mutations were accumulated in the Type 2 homeologs, supporting our prediction of less diversity of Type 1 copies. For *TTG1*, the nucleotide diversity of Type 1 across species is 0.00648, and Type 2 is 0.01053, and for *GL1*, nucleotide diversity for Type 1 is 0.009, and Type 2 is 0.013. This analysis could only be run on *TTG1* and *GL1* genes because of having the complete set of coding sequences, but not for the bHLHs. While the Type 2 copies of *TTG1* and *GL1* generally showed greater diversity, these nucleotide diversity values were not very different from each other and their nucleotide diversity patterns were quite stable within important functional motifs such as WD40 and carboxyl end of *TTG1* (Galway et al

1994, Walker et al 1999) and R2R3 MYB and TAD in *GL1* (Esch et al 1994, Stracke et al 2001, Zimmermann et al 2004).

The Ka/Ks analysis also indicated that each homeolog of each gene of interest is under purifying selection relative to the other, except for a few individual regions of apparent positive selection which are outside of known functional motifs. In addition, Tajima's test and Fu and Li D* and F* test statistics indicated that all homeologous group-wise mutations are selectively neutral controversial to the purifying selection suggested by Ka/Ks analysis. Small sample size, localized mutations, and limitations of the model assumptions could result in controversy.

The protein analyses of TTG1 showed that none of the mutations identified in the *Pachycladon* sequences have a potential negative impact on structure or function, which would suggest that both of the *TTG1* homeologs from all the study species are functional and under purifying selection as stated above.

We previously detected a 5 bp deletion in exon 1 of the *P. fastigiatum GL1* Type 2 homeolog. In the protein analysis, we observed this resulting in a translational frameshift and thereby a truncated protein of 42 amino acid residues ending in the functionally important R2 domain; the other intact GL1 proteins are ~231 amino acid residues in length. Hence, this Type 2 protein lacks three of the five essential characteristic tryptophan residues in the R2 domain, the entire R3 domain, and the transcriptional activation domain. Hence, the protein lacks the conserved 20 amino acids signature identified in the R3 MYB region needed for the interaction with the bHLHs (Zimmermann et al 2004) for the formation of the transactivation complex. The other homeolog of *GL1* (Type 1) from *P. fastigiatum* had no noticeable deleterious mutations when aligned with the most crucial regions of the *A. thaliana* GL1 protein. GO annotations obtained for these protein copies are similar to those of *A. thaliana* GL1, stating they are transcription factors functionally important in nucleic acid binding and transcriptional regulation. With an apparently functional and expressed *GL1* Type 1 copy in *P. fastigiatum*, further research should focus on identifying the molecular nature of glabrousness in this species.

Thus, our results align with our predictions that the Type 1 subgenome is in overall bias compared to the Type 2 subgenome, which may have been fractionated more.

5.3 Expression analysis of the duplicated genetic pathway

The gene expression profiles of the four key gene homeologs involved in the *TTG1-GL1-GL3*/*EGL3* complex of the trichome initiation pathway were studied in the four *Pachycladon* study species at different growth stages, depending on the material available: seedlings of *P. enysii* and *P. cheesemanii* and mature leaves of *P. fastigiatum*, *P. novae-zealandiae*, and *P. cheesemanii*. Different growth stages of different species mainly depended on their different growth rates and seedling numbers available for harvest.

The expression levels of the two homeologs of *TTG1* were the most consistent across all species and tissues tested. It was only in the mature leaves of glabrous *P. fastigiatum* that a subtle difference was detected. In all other species and tissue types, the two copies were at very similar levels. This was in keeping with the finding of low genetic diversity in both homeologs and an influence of purifying selection.

For *GL3*, the clear trend was that the Type 1 homeologs were more highly expressed than the Type 2 homeologs across nearly all species and tissue types. Remarkably, the *GL3* Type 1 homeolog was significantly expressed at a greater level than Type 2 across all samples. Whether this represents a change that occurred prior to the diversification of the group and has been retained by all species examined or has arisen independently in each species is unclear. Obtaining and analyzing complete coding region sequences for both *GL3* homeologs for all species might help clarify the issue.

Similar to *GL3*, *GL1* Type 1 homeologs clearly showed a trend of being more highly expressed than the Type 2 homeologs in four out of the five species/tissue type combinations, but a significant difference was only detected in the *P. cheesemanii* seedlings. Most of the *GL1* sequences except for *P. fastigiatum* Type 2, did not exhibit polymorphisms that were predicted to have a deleterious effect on protein structure and function. As previously described, the *P.*

fastigiatum GL1 Type 2 sequence showed a 5 bp deletion, causing a translational frameshift resulting in a truncated protein devoid of essential amino acid motifs required for the formation of the transactivation complex in the trichome initiation pathway. However, the *GL1* expression profile for *P. fastigiatum* only showed a 2-fold higher expression for the Type 1 homeolog, which was not a significant difference. It is interesting that the Type 2 copy is still expressed. The significantly greater expression of *GL1* Type 1 in *P. cheesemanii* young seedlings is interesting as no such difference was detected in the *P. cheesemanii* mature leaf samples; however, based on work in *A. thaliana*, *GL1* expression declines in older leaves. Moreover, changes in the temporal expression of *GL1* are dependent on other factors not limited to gene sequences, such as phytohormone regulation (GA) and upstream gene regulation (*GIS*) (Gan et al 2006, Lee et al 2007).

Based on the literature, overall *GL1* expression is found high in trichome precursor cells and developing trichomes in *Arabidopsis*, explaining the gene's primary known function of differentiating protodermal cells to trichomes (Larkin et al 1994). Our data reassures this theory, by showing higher expression levels for both *GL1* homeologs in young *P. cheesemanii* seedlings compared to mature *P. cheesemanii* tissue, and from that even significantly higher expression levels for Type 1 homeolog suggesting biased expression for the Type 1 subgenome.

The expression data provides some interesting insights into the retention of duplicated gene copies. More or less similar expression of both copies of *TTG1* show that both copies tend to be retained when duplicated. Of note is the highly pleiotropic nature of *TTG1*. There is evidence from the literature showing the retention of duplicated genes when they are pleiotropic and involved in multiple important pathways (Dressel & Hemleben 2009, Humphries et al 2005, Krakauer & Nowak 1999, Yan et al 2014). Moreover, when gene redundancy is not favoured in a genetic pathway, but it is required to maintain the duplicated copies of a pleiotropic gene, sub- or neo-functionalization could take place (Rastogi & Liberles 2005).

Previous studies on duplicated genomes state that genetic redundancy is sometimes observed in nascent polyploids and is reduced along the passage of time by fractionation because maintaining two copies of genes is often unfavorable for the fitness of the plant (De Smet & Van de Peer 2012, Panchy et al 2016) unless the existence of the pair of duplicates is selectively advantageous. When it is required to form a new dosage balance between the interacting partners to reduce redundancy, it is assumed that the least disruptive partners in the complex are mutated first (Birchler et al 2007), as might be the case for *GL3* in the current study. Meanwhile, redundantly expressed genes (such as *TTG1*) may over time adjust the gene dosage of the complex by developing sub- or neo-functions based on new mutations. It is predicted that a complex will ultimately regain the appropriate stoichiometry in an adjusted new gene dosage balance preferring a single subgenome (Alger & Edger 2020). In the case of *Pachycladon*, the Type 1 subgenome seems to be preferentially expressed. *Pachycladon* has already been studied for its descending dysploidy (through genome repatterning with chromosome fusions mediated by inversions, translocations, and centromere inactivation/loss) towards diploid-like genomes (n=10) since the allopolyploidization event (Mandáková et al 2010).

Overall, based on the sequence analyses, nucleotide divergence analyses, and protein analyses, Type 2 homeologs are somewhat more variable than Type 1 homeologs. Assuming a single origin of the *Pachycladon* lineage, this variation would have arisen following the hybridization/polyploidy event(s) and during subsequent diversification. Although the analyses here suggest that all homeologs are mostly under the influence of purifying selection, we note that in addition to showing great variation, the Type 2 homeologs also tend to be expressed at lower levels supporting our hypothesis predicted on more biased expression of Type 1 subgenome.

5.4 Future directions

Based on previous studies (Huang et al 2015, Joly et al 2009, Mandakova et al 2010, Zhao et al 2010) and the data presented here, the phylogenetic position of one of the *Pachycladon* progenitors seems to be within the *Brassicaceae* lineage I amongst *Arabidopsis*, *Capsella*, and *Camelina*. The position of the other progenitor is often inferred as being further down in the Brassicaceae lineage I near the *Smelowskiea* clade (Huang et al 2015, Joly et al 2009), but further resolution using more orthologs and loci is needed to infer the exact phylogenetic position of the Type 2 homeolog. Furthermore, the identity of the maternal progenitor lineage remains unresolved. The Type 2 copy lying at the base of the Lineage I was identified as the maternal copy by one group (Joly et al 2009), whereas the *Camelineae* copy was identified as the maternal copy by another group (Mandakova et al 2010). Although a recent study (Hendriks et al 2023) that examined 60 plastome genes across the family placed the maternal parent closest to *Crucihimalaya* species, which suggests that it is the Type 2 parent, the issue is somewhat unresolved. A focused study involving full chloroplast genome sequences with targeted taxa sampling needs to be completed.

Obtaining the complete sequences of bHLH homeologs will be an important next step to help build up a more complete picture of the fates of the duplicated genes. Studying intronic sequences and *cis*-regulatory sequences, particularly for *GL1* and *GL3* may contribute to a better interpretation of gene expression analysis. Targeted region sequencing such as targeted amplicon sequencing (Bybee et al 2011) or targeted enrichment sequencing (Mertes et al 2011) which provides high coverage depth for targeted regions will provide more depth of sequencing for this type of analysis. We couldn't adopt these methods in our current study because of limited time left after using traditional sequencing attempts such as Sanger sequencing and Miseq genomic sequencing.

It would be interesting to validate our homeologous gene expression data with protein-protein interactions studies using yeast-2-hybrid systems (Acharya et al 2023, Xiao et al 2022) and *A. thaliana* transgenics to functionally characterize the homeologs, benefiting from

Pachycladon's phylogenetic closeness with *A. thaliana* (Aznar-Moreno & Durrett 2017, Mei et al 2021, Tyagi et al 2019). Transgenics include mutation studies such as overexpressing *Pachycladon* highly-expressed-Type 1 homeologs in mutant *A. thaliana* strains: e.g. overexpress *Pachycladon GL3* Type 1 in *A. thaliana gl3:egl3* mutant strain. Further exploration of *P. fastigiatum GL1* duplication fate is particularly needed to further understand *GL1* homeologs' expression pattern with truncated *GL1* Type 2 copy. *P. fastigiatum* is still glabrous even with an intact *GL1* Type 1 copy. We suggested that *GL1* Type 2 copy acts as a dominant negative on *GL1* Type 1 expression to result in the glabrous phenotype. Further, to investigate if there is any role of *GL1* Type 2 copy with the trichome initiation complex, protein-protein interactions within and between types of trichome initiation complex can be tested using yeast two-hybrid system. Any interaction ideas then can be tested in mutation studies, such as to overexpress *GL1* Type 2 in *gl1* mutant *A. thaliana* to see if there are any disruptive effects on the trichome phenotype. In the meantime, *P. fastigiatum* intact *GL1* Type 1 copy can be expressed in *A. thaliana gl1* mutant background to see if it rescues the phenotype, and to draw conclusions if *GL1* Type 1 intact copy is still potent of trichome initiation. We used only three biological replicates in the current experiment owing to the limitations. In a future study, the sampling size can be increased for counting any variations that could exist in *Pachycladon* species populations from different developmental stages.

Although we aimed to study the fate of polyploidy in more genes involved in the trichome initiation pathway including the negative regulators R3 MYBs and important downstream genes of the transactivation complex such as *GL2* and *TTG2*, research-related limitations and the COVID pandemic enabled us to explore only the duplicated *TTG1-GL1-GL3*/*EGL3* transactivation complex. It would be interesting to see future studies on the role of duplicated R3 MYBs and downstream genes to illustrate a bigger picture of the fate of the entire duplicated pathway. This would potentially explain why *P.cheesemanii* seedlings start off bald but later developed branched trichomes. Likewise, there is enormous potential to grow this study directing multiple areas of unresolved questions.

REFERENCES

IDT OligoAnalyzer. Integrated DNA Technologies

IDT UNAFold. Integrated DNA Technologies

Acharya BR, Zhao C, Reyes LAR, Ferreira JF, Sandhu D. 2023. Understanding the salt overly sensitive pathway in Prunus: Identification and characterization of NHX, CIPK, and CBL genes. *The Plant Genome*: e20371

Adams K, Wendel J. 2004. Exploring the genomic mysteries of polyploidy in cotton. *Biological Journal of the Linnean Society* 82: 573-81

Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168: 2217-26

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8: 135-41

Afrin S, Nuruzzaman M, Zhu J, Luo Z-Y. *2014*.

Ahrberg CD, Neužil P. 2015. Doubling Throughput of a Real-Time PCR. *Scientific Reports* 5: 12595

Airoldi CA, Hearn TJ, Brockington SF, Webb AAR, Glover BJ. 2019. TTG1 proteins regulate circadian activity as well as epidermal cell fate and pigmentation. *Nat Plants* 5: 1145-53

Al-Shehbaz IA, Beilstein MA, Kellogg EA. 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Systematics and Evolution* 259: 89-120

Al-Shehbaz IA, O'Kane SL, Price RA. 1999. Generic Placement of Species Excluded from Arabidopsis (Brassicaceae). *Novon* 9: 296-307

Alfandary R. 2015. 4 Tips for Efficient Primer Design.

Alger EI, Edger PP. 2020. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Current opinion in plant biology* 54: 108-13

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215: 403-10

Andersen CL, Jensen JL, Ørntoft TF. 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer research* 64: 5245-50

Arezi B, Xing W, Sorge JA, Hogrefe HH. 2003. Amplification efficiency of thermostable DNA polymerases. *Anal Biochem* 321: 226-35

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. *Current opinion in plant biology* 15: 140-46

Ascencio D, Diss G, Gagnon-Arsenault I, Dubé AK, DeLuna A, Landry CR. 2021. Expression attenuation as a mechanism of robustness against gene duplication. *Proceedings of the National Academy of Sciences* 118: e2014345118

Atchley WR, Fitch WM. 1997. A natural classification of the basic helix-loop-helix class of transcription factors. *Proceedings of the National Academy of Sciences* 94: 5172-76

Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444: 171-78

Aznar-Moreno J, Durrett T. 2017. Simultaneous targeting of multiple gene homeologs to alter seed oil production in Camelina sativa. *Plant and Cell Physiology* 58: 1260-67

Bae H, Coller J. 2022. Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. *Molecular cell* 82: 1467-76

Balkunde R, Pesch M, Hülskamp M. 2010. Trichome patterning in Arabidopsis thaliana: from genetic to molecular models. *Current topics in developmental biology* 91: 299-321

Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B. 2011. Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. *New Phytol* 192: 760-74

Behling AH, Shepherd LD, Cox MP. 2020. The importance and prevalence of allopolyploidy in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand* 50: 189-210

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American journal of botany* 93: 607-19
Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. *The Plant Cell* 23: 1719-28
Bennett MD. 2004. Perspectives on polyploidy in plants–ancient and neo. *Biological Journal of the Linnean Society* 82: 411-23
Bernhardt C, Lee MM, Gonzalez A, Zhang F, Lloyd A, Schiefelbein J. 2003. The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root. *Development* 130: 6431-39
Bicknell RA, Heenan PB, Dawson MI, Fletcher PJ, Christie MC. 2009. Matromorphy in Pachycladon exile (brassicaceae) revealed by interspecific hybridisation. *New Zealand Journal of Botany* 47: 139-48
Birchler JA. 2012. Genetic consequences of polyploidy in plants In *Polyploidy and genome evolution*, pp. 21-32: Springer
Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol* 234: 275-88
Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist* 186: 54-62
Birchler JA, Yao H, Chudalayandi S. 2007. Biological consequences of dosage dependent gene regulatory systems. *Biochim Biophys Acta* 1769: 422-8
Bird KA, VanBuren R, Puzey JR, Edger PP. 2018. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytologist* 220: 87-93
Blakeslee AF. 1941. Effect of induced polyploidy in plants. *The American Naturalist* 75: 117-35
Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* 13: 137-44
Blanc G, Wolfe KH. 2004a. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16: 1679-91
Blanc G, Wolfe KH. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* 16: 1667-78
Bloomer RH, Juenger TE, Symonds VV. 2012. Natural variation in GL1 and its effects on trichome density in Arabidopsis thaliana. *Molecular Ecology* 21: 3501-15
Bomblies K, Doebley JF. 2006. Pleiotropic effects of the duplicate maize FLORICAULA/LEAFY genes zfl1 and zfl2 on traits under selection during maize domestication. *Genetics* 172: 519-31
Bottani S, Zabet NR, Wendel JF, Veitia RA. 2018. Gene expression dominance in allopolyploids: hypotheses and models. *Trends in Plant Science* 23: 393-402
Bouyer D, Geier F, Kragler F, Schnittger A, Pesch M, et al. 2008. Two-dimensional patterning by a trapping/depletion mechanism: the role of TTG1 and GL3 in Arabidopsis trichome formation. *PLoS biology* 6: e141
Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433-38
Bowman JL, Briginshaw LN, Fisher TJ, Flores-Sandoval E. 2019. Something ancient and something neofunctionalized—evolution of land plant hormone signaling pathways. *Current Opinion in Plant Biology* 47: 64-72
Bridges CB. 1936. The bar" gene" a duplication. *Science* 83: 210-11
Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS. 2006. Unequal genetic redundancies in Arabidopsis--a neglected phenomenon? *Trends Plant Sci* 11: 492-8
Buggs RJ, Chamala S, Wu W, Gao L, May GD, et al. 2010. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular ecology* 19: 132-46

Buggs RJ, Chamala S, Wu W, Tate JA, Schnable PS, et al. 2012. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr Biol* 22: 248-52

Buggs RJ, Wendel JF, Doyle JJ, Soltis DE, Soltis PS, Coate JE. 2014. The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130354

Buggs RJ, Zhang L, Miles N, Tate JA, Gao L, et al. 2011. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr Biol* 21: 551-6

Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, et al. 2021. Gradual evolution of allopolyploidy in Arabidopsis suecica. *Nature ecology & evolution* 5: 1367-81

Bustin S, Huggett J. 2017. qPCR primer design revisited. *Biomolecular Detection and Quantification* 14: 19-28

Bustin SA. 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* 25: 169-93

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, et al. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55: 611-22

Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, et al. 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome biology and evolution* 3: 1312-23

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics* 23: 1875-82

Carretero-Paulet L, Fares MA. 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular biology and evolution* 29: 3541-51

Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martínez-García JF, Bilbao-Castro JR, Robertson DL. 2010. Genome-Wide Classification and Evolutionary Analysis of the bHLH Family of Transcription Factors in Arabidopsis, Poplar, Rice, Moss, and Algae *Plant Physiology* 153: 1398-412

Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV. 2010. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biology* 11: R125

Chapman BA, Bowers JE, Feltus FA, Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences* 103: 2730-35

Chapman MA, Burke JM. 2007. GENETIC DIVERGENCE AND HYBRID SPECIATION. *Evolution* 61: 1773-80

Chen G, Li X, Qiao X, Li J, Wang L, et al. 2019. Genome-wide survey and expression analysis of the SLAC/SLAH gene family in pear (Pyrus bretschneideri) and other members of the Rosaceae. *Genomics* 111: 1097-107

Chen H-W, Bandyopadhyay S, Shasha DE, Birnbaum KD. 2010. Predicting genome-wide redundancy using machine learning. *BMC Evol Biol* 10: 357

Chen S, Wang S. 2019. GLABRA2, a Common Regulator for Epidermal Cell Fate Determination and Anthocyanin Biosynthesis in Arabidopsis. *International Journal of Molecular Sciences* 20: 4997

Chen ZJ. 2007. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual Review of Plant Biology* 58: 377-406

Chen ZJ. 2010. Molecular mechanisms of polyploidy and hybrid vigor. *Trends in plant science* 15: 57-71

Chen ZJ, Ni Z. 2006. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays* 28: 240-52

Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, et al. 2016. Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. *New Phytologist* 211: 288-99

Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* 4: 258-68
Cheng F, Wu J, Fang L, Sun S, Liu B, et al. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PloS one* 7: e36442
Cheng Z, Sun L, Qi T, Zhang B, Peng W, et al. 2011. The bHLH transcription factor MYC3 interacts with the Jasmonate ZIM-domain proteins to mediate jasmonate response in Arabidopsis. *Mol Plant* 4: 279-88
Chezem WR, Clay NK. 2016. Regulation of plant secondary metabolism and associated specialized cell development by MYBs and bHLHs. *Phytochemistry* 131: 26-43
Chu Y, Corey DR. 2012. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther* 22: 271-4
Clark JW, Donoghue PC. 2018. Whole-genome duplication and plant macroevolution. *Trends in plant science* 23: 933-45
Colle M, Leisner CP, Wai CM, Ou S, Bird KA, et al. 2019. Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience* 8: giz012
Comai L. 2000. Genetic and epigenetic interactions in allopolyploid plants. *Plant Gene Silencing*: 267-79
Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* 6: 836-46
Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* 19: 91-98
Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* 9: 938-50
Coudert E, Gehant S, de Castro E, Pozzato M, Baratin D, et al. 2023. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* 39
Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome research* 16: 738-49
Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR. 2005. Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol* 139: 5-17
D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488: 213-17
Dayhoff MO, Eck RV. 1972. *Atlas of protein sequence and structure*. National Biomedical Research Foundation.
De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20: 591-7
De Smet R, Sabaghian E, Li Z, Saeys Y, Van de Peer Y. 2017. Coordinated Functional Divergence of Genes after Genome Duplication in Arabidopsis thaliana *The Plant Cell* 29: 2786-800
De Smet R, Van de Peer Y. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current opinion in plant biology* 15: 168-76
De Vries H. 1910. *The mutation theory: Experiments and observations on the origin of species in the vegetable kingdom*. Kegan Paul.
Dekkers BJW, Willems L, Bassel GW, van Bolderen-Veldkamp RP, Ligterink W, et al. 2011. Identification of Reference Genes for RT–qPCR Expression Analysis in Arabidopsis and Tomato Seeds. *Plant and Cell Physiology* 53: 28-37
Del Pozo JC, Ramirez-Parra E. 2015. Whole genome duplications in plants: an overview from Arabidopsis. *Journal of Experimental Botany* 66: 6991-7003
Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research* 12: 1075-79
DeVries H. 1915. The coefficient of mutation in Oenothera biennis L. *Botanical Gazette* 59: 169-96

Díaz-Pérez A, López-Álvarez D, Sancho R, Catalán P. 2018. Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus Brachypodium using minimum evolution, coalescence and maximum likelihood approaches. *Molecular Phylogenetics and Evolution* 127: 256-71

Dodsworth S, Chase MW, Leitch AR. 2016. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society* 180: 1-5

Dombrecht B, Xue GP, Sprague SJ, Kirkegaard JA, Ross JJ, et al. 2007. MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. *Plant Cell* 19: 2225-45

Dong Y, Gupta S, Sievers R, Wargent JJ, Wheeler D, et al. 2019. Genome draft of the Arabidopsis relative Pachycladon cheesemanii reveals novel strategies to tolerate New Zealand's high ultraviolet B radiation environment. *BMC Genomics* 20: 838

Doroshkov AV, Konstantinov DK, Afonnikov DA, Gunbin KV. 2019. The evolution of gene regulatory networks controlling Arabidopsis thaliana L. trichome development. *BMC Plant Biology* 19: 53

Doyle JJ, Coate JE. 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences* 180: 1-52

Doyle JJ, Doyle JL. 1990. Isolation ofplant DNA from fresh tissue. *Focus* 12: 39-40

Doyle JJ, Sherman-Broyles S. 2017. Double trouble: taxonomy and definitions of polyploidy. *New Phytologist* 213: 487-93

Dressel A, Hemleben V. 2009. Transparent Testa Glabra 1 (TTG1) and TTG1-like genes in Matthiola incana R. Br. and related Brassicaceae and mutation in the WD-40 motif. *Plant Biology* 11: 204-12

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L. 2010. MYB transcription factors in Arabidopsis. *Trends in plant science* 15: 573-81

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* 17: 699-717

Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, et al. 2019. Origin and evolution of the octoploid strawberry genome. *Nature Genetics* 51: 541-47

Esch JJ, Oppenheimer DG, Marks MD. 1994. Characterization of a weak allele of the GL1 gene of Arabidopsis thaliana. *Plant Molecular Biology* 24: 203-07

Expósito-Rodríguez M, Borges AA, Borges-Pérez A, Pérez JA. 2008. Selection of internal control genes for quantitative real-time RT-PCR studies during tomato development process. *BMC plant biology* 8: 131-31

Farhat P, Hidalgo O, Robert T, Siljak-Yakovlev S, Leitch IJ, et al. 2019. Polyploidy in the Conifer Genus Juniperus: An Unexpectedly High Rate. *Front Plant Sci* 10: 676-76

Fay JC, Wu CI. 2001. The neutral theory in the genomic era. *Curr Opin Genet Dev* 11: 642-6

Ferreira de Carvalho J, Lucas J, Deniot G, Falentin C, Filangi O, et al. 2019. Cytonuclear interactions remain stable during allopolyploid evolution despite repeated whole-genome duplications in Brassica. *Plant J* 98: 434-47

Flagel L, Udall J, Nettleton D, Wendel J. 2008. Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC biology* 6: 1-9

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytol* 183: 557-64

Flatschacher D, Speckbacher V, Zeilinger S. 2022. qRAT: an R-based stand-alone application for relative expression analysis of RT-qPCR data. *BMC Bioinformatics* 23: 286

Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML. 2007. Evolutionary dynamics of Waxy and the origin of hexaploid Spartina species (Poaceae). *Mol Phylogenet Evol* 43: 1040-55

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn* 4: 25-40

Freeling M. 2009. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology* 60: 433-53
Freeling M, Scanlon MJ, Fowler JE. 2015. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr Opin Genet Dev* 35: 110-8
Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693-709
Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. *Plant Cell* 19: 3403-17
Galway ME, Masucci JD, Lloyd AM, Walbot V, Davis RW, Schiefelbein JW. 1994. The TTG gene is required to specify epidermal cell fate and cell patterning in the Arabidopsis root. *Developmental biology* 166: 740-54
Gan Y, Kumimoto R, Liu C, Ratcliffe O, Yu H, Broun P. 2006. GLABROUS INFLORESCENCE STEMS modulates the regulation by gibberellins of epidermal differentiation and shoot maturation in Arabidopsis. *The Plant Cell* 18: 1383-95
Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in Arabidopsis. *Molecular biology and evolution* 24: 2298-309
Garnock-Jones PJ. 1991. Gender dimorphism in Cheesemania wallii (Brassicaceae). *New Zealand Journal of Botany* 29: 87-90
Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two evolutionarily distinct classes of paleopolyploidy. *Molecular biology and evolution* 31: 448-54
Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC. 2000. Maize as a model for the evolution of plant nuclear genomes. *Proceedings of the National Academy of Sciences* 97: 7008-15
Geneious 9.1.8. 2017.
GenScript. 2021. Sanger Sequencing vs. Next-Generation Sequencing (NGS).
German DA, Friesen N, Neuffer B, Al-Shehbaz IA, Hurka H. 2009. Contribution to ITS phylogeny of the Brassicaceae, with special reference to some Asian taxa. *Plant Systematics and Evolution* 283: 33-56
Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome research* 17: 669-81
Gerstel D, Phillips L. 1957. Segregation in new allopolyploids of Gossypium. II. Tetraploid combinations. *Genetics* 42: 783
Ginzinger DG. 2002. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. *Experimental hematology* 30: 503-12
Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, et al. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *Rna* 16: 991-1006
Glover BJ, Bunnewell S, Martin C. 2004. Convergent evolution within the genus Solanum: the specialised anther cone develops through alternative pathways. *Gene* 331: 1-7
Glover BJ, Perez-Rodriguez M, Martin C. 1998. Development of several epidermal cell types can be specified by the same MYB-related plant transcription factor. *Development* 125: 3497-508
Gong L, Olson M, Wendel JF. 2014. Cytonuclear Evolution of Rubisco in Four Allopolyploid Lineages. *Molecular Biology and Evolution* 31: 2624-36
Gong L, Salmon A, Yoo M-J, Grupp KK, Wang Z, et al. 2012. The Cytonuclear Dimension of Allopolyploid Evolution: An Example from Cotton Using Rubisco. *Molecular Biology and Evolution* 29: 3023-36
Gonzalez A, Mendenhall J, Huo Y, Lloyd A. 2009. TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. *Developmental biology* 325: 412-21
Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17: 333-51
Grant V. 1981. Plant speciation, Columbia University Press

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *science* 185: 862-64
Gray VE, Hause RJ, Fowler DM. 2017. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* 207: 53-61
Grebe M. 2012. The patterning of epidermal hairs in Arabidopsis--updated. *Curr Opin Plant Biol* 15: 31-7
Grover C, Gallagher J, Szadkowski E, Yoo M, Flagel L, Wendel J. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* 196: 966-71
Grover CE, Gallagher JP, Jareczek JJ, Page JT, Udall JA, et al. 2015. Re-evaluating the phylogeny of allopolyploid Gossypium L. *Molecular Phylogenetics and Evolution* 92: 45-52
Guo H, Lee T-H, Wang X, Paterson AH. 2013. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. *Plant Physiology* 162: 769-78
Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, et al. 2006. Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theoretical and Applied Genetics* 113: 831-45
Gurson N. 2015. When Do I Use Sanger Sequencing vs NGS? : ThermoFisher Scientific
Ha M, Li WH, Chen ZJ. 2007. External factors accelerate expression divergence between duplicate genes. *Trends Genet* 23: 162-6
Hahn F, Eisenhut M, Mantegazza O, Weber AP. 2018. Homology-directed repair of a defective glabrous gene in Arabidopsis with Cas9-based gene targeting. *Front Plant Sci* 9: 424
Han G, Li Y, Yang Z, Wang C, Zhang Y, Wang B. 2022. Molecular mechanisms of plant trichome development. *Front Plant Sci* 13: 910228
Hao Y, Zong X, Ren P, Qian Y, Fu A. 2021. Basic Helix-Loop-Helix (bHLH) transcription factors regulate a wide range of functions in Arabidopsis. *International Journal of Molecular Sciences* 22: 7152
Hauser M-T. 2014. Molecular basis of natural variation and environmental control of trichome patterning. *Front Plant Sci* 5
Hauser M-T, Harr B, Schlötterer 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: 1754-63
Heenan P, Mitchell A, Koch M. 2002. Molecular systematics of the New Zealand Pachycladon (Brassicaceae) complex: generic circumscription and relationships to Arabidopsis sens. lat. and Arabis sens. lat. *New Zealand Journal of Botany* 40: 543-62
Heenan PB. 1999. Artificial intergeneric hybrids between the New Zealand endemic Ischnocarpus and Pachycladon (Brassicaceae). *New Zealand Journal of Botany* 37: 595-601
Heenan PB, Dawson MI, Smissen RD, Bicknell RA. 2008. An artificial intergeneric hybrid derived from sexual hybridization between the distantly related Arabidopsis thaliana and Pachycladon cheesemanii (Brassicaceae). *Botanical Journal of the Linnean Society* 157: 533-44
Heenan PB, Mitchell AD. 2003. Phylogeny, biogeography and adaptive radiation of *Pachycladon* (Brassicaceae) in the mountains of South Island, New Zealand. *Journal of Biogeography* 30: 1737–49
Hegarty MJ, Barker GL, Wilson ID, Abbott RJ, Edwards KJ, Hiscock SJ. 2006. Transcriptome Shock after Interspecific Hybridization in Senecio Is Ameliorated by Genome Duplication. *Current Biology* 16: 1652-59
Hellemans J, Vandesompele J. 2011. qPCR data analysis-unlocking the secret to successful results. *PCR troubleshooting and optimization: the essential guide* 1: 13
Hendriks KP, Kiefer C, Al-Shehbaz IA, Bailey CD, van Huysduynen AH, et al. 2023. Global Brassicaceae phylogeny based on filtering of 1,000-gene dataset. *Current Biology*
Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89: 10915-19

Higuchi R, Fockler C, Dollinger G, Watson R. 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology* 11: 1026-30

Hruz T, Wyss M, Docquier M, Pfaffl MW, Masanetz S, et al. 2011. RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* 12: 156

Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: An overview. *Human Immunology* 82: 801-11

Hu X, Hao C, Cheng Z-M, Zhong Y. 2019. Genome-Wide Identification, Characterization, and Expression Analysis of the Grapevine Superoxide Dismutase (SOD) Family. *International Journal of Genomics* 2019: 13

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, et al. 2001. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* 29: 102-5

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, et al. 2015. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. *Molecular Biology and Evolution* 33: 394-412

Hufton AL, Panopoulou G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Current opinion in genetics & development* 19: 600-06

Hughes TE, Langdale JA, Kelly S. 2014. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome research* 24: 1348-55

Hülskamp M, Misŕa S, Jürgens G. 1994. Genetic dissection of trichome cell development in Arabidopsis. *Cell* 76: 555-66

Humphries JA, Walker AR, Timmis JN, Orford SJ. 2005. Two WD-repeat genes from cotton are functional homologues of the Arabidopsis thaliana TRANSPARENT TESTA GLABRA1 (TTG1) gene. *Plant molecular biology* 57: 67-81

Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* 18: 486

Illumina. 2022. Key differences between next-generation sequencing and Sanger sequencing-Understanding when NGS can be a more effective option.

Ishida T, Hattori S, Sano R, Inoue K, Shirano Y, et al. 2007. Arabidopsis TRANSPARENT TESTA GLABRA2 is directly regulated by R2R3 MYB transcription factors and is involved in regulation of GLABRA2 transcription in epidermal differentiation. *Plant Cell* 19: 2531-43

Ishida T, Kurata T, Okada K, Wada T. 2008. A genetic regulatory network in the development of trichomes and root hairs. *Annu. Rev. Plant Biol.* 59: 365-86

Jackson S, Chen ZJ. 2010. Genomic and expression plasticity of polyploidy. *Current opinion in plant biology* 13: 153-59

Jenczewski E, Chèvre A-M, Alix K. 2013. Chromosomal and gene expression changes in Brassica allopolyploids. *Polyploid and hybrid genomics*: 171-86

Jiang W-k, Liu Y-l, Xia E-h, Gao L-z. 2013. Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants *Plant Physiology* 161: 1844-61

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13: R3

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97

Jin Y, Liu F, Huang W, Sun Q, Huang X. 2019. Identification of reliable reference genes for qRT-PCR in the ephemeral plant Arabidopsis pumila based on full-length transcriptome data. *Scientific Reports* 9: 8408

Joly S, Heenan PB, Lockhart PJ. 2009. A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of Pachycladon (Brassicaceae) in New Zealand. *Mol Phylogenet Evol* 51: 365-72

Joly S, Heenan PB, Lockhart PJ. 2013. Species Radiation by Niche Shifts in New Zealand's Rockcresses (*Pachycladon*, Brassicaceae). *Syst. Biol.* 63: 192–202
Joly S, Heenan PB, Lockhart PJ. 2014. Species Radiation by Niche Shifts in New Zealand's Rockcresses (Pachycladon, Brassicaceae). *Systematic Biology* 63: 192-202
Joly S, Rauscher JT, Sherman-Broyles SL, Brown AH, Doyle JJ. 2004. Evolutionary dynamics and preferential expression of homeologous 18S-5.8S-26S nuclear ribosomal genes in natural and artificial glycine allopolyploids. *Mol Biol Evol* 21: 1409-21
Jones DM, Vandepoele K. 2020. Identification and evolution of gene regulatory networks: insights from comparative studies in plants. *Current opinion in plant biology* 54: 42-48
Joshi P, Ansari H, Dickson R, Ellison NW, Skema C, Tate JA. 2022. Polyploidy on islands – concerted evolution and gene loss amid chromosomal stasis. *Annals of Botany* 131: 33-44
Kaltenegger E, Leng S, Heyl A. 2018. The effects of repeated whole genome duplication events on the evolution of cytokinin signaling pathway. *BMC Evol Biol* 18: 76
Karabourniotis G, Liakopoulos G, Nikolopoulos D, Bresta P. 2020. Protective and defensive roles of non-glandular trichomes against multiple stresses: structure–function coordination. *Journal of Forestry Research* 31: 1-12
Kawagoe T, Shimizu KK, Kakutani T, Kudoh H. 2011. Coexistence of trichome variation in a natural plant population: a combined study using ecological and candidate gene approaches. *PloS one* 6: e22184
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* 10: 845-58
Kihara H, Ono T. 1926. Chromosomenzahlen und systematische gruppierung der Rumex-arten. *Zeitschrift für Zellforschung und mikroskopische Anatomie* 4: 475-81
Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.
Kirik V, Lee MM, Wester K, Herrmann U, Zheng Z, et al. 2005. Functional diversification of MYB23 and GL1 genes in trichome morphogenesis and initiation. *Development* 132: 1749-49
Kirik V, Schnittger A, Radchuk V, Adler K, Hülskamp M, Bäumlein H. 2001. Ectopic expression of the Arabidopsis AtMYB23 gene induces differentiation of trichome cells. *Dev Biol* 235: 366-77
Kirik V, Simon M, Huelskamp M, Schiefelbein J. 2004a. The ENHANCER OF TRY AND CPC1 gene acts redundantly with TRIPTYCHON and CAPRICE in trichome and root hair cell patterning in Arabidopsis. *Dev Biol* 268: 506-13
Kirik V, Simon M, Wester K, Schiefelbein J, Hulskamp M. 2004b. ENHANCER of TRY and CPC 2 (ETC2) reveals redundancy in the region-specific control of trichome development of Arabidopsis. *Plant Mol Biol* 55: 389-98
Kliebenstein DJ. 2008. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS one* 3: e1838
Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* 279: 5048-57
Kontanis EJ, Reed FA. 2006. Evaluation of real-time PCR amplification efficiencies to detect PCR inhibitors. *Journal of forensic sciences* 51: 795-804
Krakauer DC, Nowak MA. 1999. Evolutionary preservation of redundant duplicated genes. *Seminars in Cell & Developmental Biology* 10: 555-59
Kryvych S, Kleessen S, Ebert B, Kersten B, Fisahn J. 2011. Proteomics–the key to understanding systems biology of Arabidopsis trichomes. *Phytochemistry* 72: 1061-70
Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, et al. 2006. The real-time polymerase chain reaction. *Molecular Aspects of Medicine* 27: 95-125
Kumar A, Kumar S, Venkatesh K, Singh NK, Mandal PK, Sinha SK. 2022. Physio-molecular traits of contrasting bread wheat genotypes associated with 15N influx exhibiting homeolog expression bias in nitrate transporter genes under different external nitrate concentrations. *Planta* 255: 104

Kurata T, Ishida T, Kawabata-Awai C, Noguchi M, Hattori S, et al. 2005. Cell-to-cell movement of the CAPRICE protein in Arabidopsis root epidermal cell differentiation. *Development* 132: 5387-98

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. 2004. Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics* 166: 935-45

Larkin JC, Oppenheimer DG, Lloyd AM, Paparozzi ET, Marks MD. 1994. Roles of the GLABROUS1 and TRANSPARENT TESTA GLABRA Genes in Arabidopsis Trichome Development. *The Plant cell* 6: 1065-76

Larkin JC, Oppenheimer DG, Pollock S, Marks MD. 1993. Arabidopsis GLABROUS1 Gene Requires Downstream Sequences for Function. *Plant Cell* 5: 1739-48

Larkin JC, Walker JD, Bolognesi-Winfield AC, Gray JC, Walker AR. 1999. Allele-specific interactions between ttg and gl1 during trichome development in Arabidopsis thaliana. *Genetics* 151: 1591-604

Larkin JC, Young N, Prigge M, Marks MD. 1996. The control of trichome spacing and number in Arabidopsis. *Development* 122: 997-1005

Láruson ÁJ, Yeaman S, Lotterhos KE. 2020. The importance of genetic redundancy in evolution. *Trends in ecology & evolution* 35: 809-22

Lee JJ, Woodward AW, Chen ZJ. 2007. Gene Expression Changes and Early Events in Cotton Fibre Development. *Annals of Botany* 100: 1391-401

Lee MM, Schiefelbein J. 1999. WEREWOLF, a MYB-related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning. *Cell* 99: 473-83

Lee MM, Schiefelbein J. 2001. Developmentally distinct MYB genes encode functionally equivalent proteins in Arabidopsis. *Development* 128: 1539-46

Lee S, Choi S, Jeon D, Kang Y, Kim C. 2020. Evolutionary impact of whole genome duplication in Poaceae family. *Journal of Crop Science and Biotechnology* 23: 413-25

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* 6: 1110-16

Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of cis-and trans-effects on gene expression. *Proceedings of the National Academy of Sciences* 105: 14471-76

Lemos B, Meiklejohn CD, Hartl DL. 2004. Regulatory evolution across the protein interaction network. *Nat Genet* 36: 1059-60

Levine MT, Holloway AK, Arshad U, Begun DJ. 2007. Pervasive and largely lineage-specific adaptive protein evolution in the dosage compensation complex of Drosophila melanogaster. *Genetics* 177: 1959-62

Levy AA, Feldman M. 2004. Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization. *Biological journal of the Linnean society* 82: 607-13

Li J, Zhang Z, Vang S, Yu J, Wong GK-S, Wang J. 2009a. Correlation Between Ka/Ks and Ks is Related to Substitution Model and Evolutionary Lineage. *Journal of Molecular Evolution* 68: 414-23

Li L, Briskine R, Schaefer R, Schnable PS, Myers CL, et al. 2016. Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC genomics* 17: 1-15

Li Q, Qiao X, Yin H, Zhou Y, Dong H, et al. 2019. Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). *Horticulture research* 6

Li SF, Allen PJ, Napoli RS, Browne RG, Pham H, Parish RW. 2020. MYB–bHLH–TTG1 regulates Arabidopsis seed coat biosynthesis pathways directly and indirectly via multiple tiers of transcription factors. *Plant and Cell Physiology* 61: 1005-18

Li SF, Milliken ON, Pham H, Seyit R, Napoli R, et al. 2009b. The Arabidopsis MYB5 transcription factor regulates mucilage synthesis, seed coat development, and trichome morphogenesis. *Plant Cell* 21: 72-89

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *TRENDS in Genetics* 21: 602-07

Li X, Noll M. 1994. Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions. *Nature* 367: 83-87
Li Z, McKibben MT, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and processes of diploidization in land plants. *Annual Review of Plant Biology* 72: 387-410
Liang G, He H, Li Y, Ai Q, Yu D. 2014. MYB82 functions in regulation of trichome development in Arabidopsis. *Journal of experimental botany* 65: 3215-23
Liang Z, Schnable JC. 2018. Functional divergence between subgenomes and gene pairs after whole genome duplications. *Molecular plant* 11: 388-97
Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, et al. 2008. Rapid chromosome evolution in recently formed polyploids in Tragopogon (Asteraceae). *PloS one* 3: e3353
Lin Q, Aoyama T. 2012. Pathways for epidermal cell differentiation via the homeobox gene GLABRA2: update on the roles of the classic regulator. *J Integr Plant Biol* 54: 729-37
Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol* 29: 365-79
Liu B, Xu C, Zhao N, Qi B, Kimatu JN, et al. 2009. Rapid genomic changes in polyploid wheat and related species: implications for genome evolution and genetic improvement. *Journal of Genetics and Genomics* 36: 519-28
Liu C, Zhang TZ. 2019. Functional diversifications of GhERF1 duplicate genes after the formation of allotetraploid cotton. *J Integr Plant Biol* 61: 60-74
Liu L, Li Y, Li S, Hu N, He Y, et al. 2012. Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012: 251364
Liu X-L, Jiang F-F, Wang Z-W, Li X-Y, Li Z, et al. 2017. Wider geographic distribution and higher diversity of hexaploids than tetraploids in Carassius species complex reveal recurrent polyploidy effects on adaptive evolution. *Scientific Reports* 7: 5395
Lloyd AM, Walbot V, Davis RW. 1992. Arabidopsis and Nicotiana anthocyanin production activated by maize regulators R and C1. *Science* 258: 1773-75
Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* 14: 265-79
Lu J, Li J-N, Lei B, Wang S-G, Chai Y-R. 2009. Molecular cloning and characterization of twoBrassica napus TTG1 genes reveal genus-specific nucleotide preference, extreme protein-level conservation and fast divergence of organ-specificity. *Genes & Genomics* 31: 129-42
Lu Y, Rausher MD. 2003. Evolutionary Rate Variation in Anthocyanin Pathway Genes. *Molecular Biology and Evolution* 20: 1844-53
Luo X, Hu Q, Zhou P, Zhang D, Wang Q, et al. 2017. Chasing ghosts: allopolyploid origin of Oxyria sinensis (Polygonaceae) from its only diploid congener and an unknown ancestor. *Molecular Ecology* 26: 3037-49
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-5
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459-73
Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* 159: 1789-804
Mable B. 2004. 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society* 82: 453-66
Madlung A. 2013. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* 110: 99-104
Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, et al. 2005. Genomic changes in synthetic Arabidopsis polyploids. *The Plant Journal* 41: 221-30
Maes L, Inzé D, Goossens A. 2008. Functional Specialization of the TRANSPARENT TESTA GLABRA1 Network Allows Differential Hormonal Control of Laminal and Marginal Trichome Initiation in Arabidopsis Rosette Leaves. *Plant Physiology* 148: 1453-64
Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. *Journal of Genetics* 92: 155-61

Makałowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences* 95: 9407-12
Manda´kova´ T, Joly S, Krzywinski M, Mummenhoff K, Lysaka MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *The Plant Cell* 22: 2277–90
Mandáková T, Heenan PB, Lysak MA. 2010. Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evol Biol* 10: 367-67
Mandakova T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of Arabidopsis. *Plant Cell* 22: 2277-90
Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal* 91: 3-21
Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus Viola (Violaceae). *Systematic biology* 64: 84-101
Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133-41
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-80
Markus Lange B, Turner GW. 2013. Terpenoid biosynthesis in trichomes—current status and future opportunities. *Plant biotechnology journal* 11: 2-22
Martin C, Bhatt K, Baumann K, Jin H, Zachgo S, et al. 2002. The mechanics of cell fate determination in petals. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357: 809-13
McBreen K, Heenan PB. 2006. Phylogenetic relationships of Pachycladon (Brassicaceae) species based on three nuclear and two chloroplast DNA markers. *New Zealand Journal of Botany* 44: 377-86
McClintock B. 1929. A cytological and genetical study of triploid maize. *Genetics* 14: 180
McCollum GD. 1957. Comparative studies of chromosome pairing in natural and induced tetraploid Dactylis. *Chromosoma* 9: 571-605
McGrath C, Lynch M. 2012. Evolutionary significance of whole-genome duplication In *Polyploidy and genome evolution*, pp. 1-20: Springer
Mei F, Chen B, Li F, Zhang Y, Kang Z, et al. 2021. Overexpression of the wheat NAC transcription factor TaSNAC4-3A gene confers drought tolerance in transgenic Arabidopsis. *Plant Physiology and Biochemistry* 160: 37-50
Méndez-Vigo B, Ausín I, Zhu W, Mollá-Morales A, Balasubramanian S, Alonso-Blanco C. 2019. Genetic Interactions and Molecular Evolution of the Duplicated Genes ICARUS2 and ICARUS1 Help Arabidopsis Plants Adapt to Different Ambient Temperatures. *The Plant Cell* 31: 1222-37
Meng Y, Lyu X, Liu J, Gao W, Ma Y, et al. 2023. Structural variation of GL1 gene determines the trichome formation in Brassica juncea. *Theoretical and Applied Genetics* 136: 75
Mershon JP, Becker M, Bickford CP. 2015. Linkage between trichome morphology and leaf optical properties in New Zealand alpine Pachycladon (Brassicaceae). *New Zealand Journal of Botany* 53: 175-82
Mertes F, ElSharawy A, Sauer S, van Helvoort JM, Van Der Zaag P, et al. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in functional genomics* 10: 374-86
Metzker ML. 2005. Emerging technologies in DNA sequencing. *Genome Res* 15: 1767-76
Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46
Mirzaei M, Pascovici D, Keighley T, George I, Voelckel C, et al. 2011. Shotgun proteomic profiling of five species of New Zealand Pachycladon. *Proteomics* 11: 166-71
Mitchell AD, Heenan PB. 2000. Systematic Relationships of New Zealand Endemic Brassicaceae Inferred from nrDNA ITS Sequence Data. *Systematic Botany* 25: 98-105

Mitchell AD, Heenan PB. 2002. Genetic variation within the Pachycladon (Brassicaceae) complex based on fluorescent AFLP data. *Journal of the Royal Society of New Zealand* 32: 427-43
Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, et al. 2014. Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish Raphanus raphanistrum and three other Brassicaceae species. *The Plant Cell* 26: 1925-37
Moghe GD, Shiu SH. 2014. The causes and molecular consequences of polyploidy in flowering plants. *Ann N Y Acad Sci* 1320: 16-34
Moreno-Romero J, Del Toro-De León G, Yadav VK, Santos-González J, Köhler C. 2019. Epigenetic signatures associated with imprinted paternally expressed genes in the Arabidopsis endosperm. *Genome biology* 20: 1-11
Morohashi K, Grotewold E. 2009. A Systems Approach Reveals Regulatory Circuitry for Arabidopsis Trichome Initiation by the GL3 and GL1 Selectors. *PLOS Genetics* 5: e1000396
Morohashi K, Zhao M, Yang M, Read B, Lloyd A, et al. 2007. Participation of the Arabidopsis bHLH Factor GL3 in Trichome Initiation Regulatory Events. *Plant Physiology* 145: 736-46
Morrison J, Rajhathy T. 1960. Frequency of quadrivalents in autotetraploid plants. *Nature* 187: 528-30
Muller HJ. 1925. Why Polyploidy is Rarer in Animals Than in Plants. *The American Naturalist* 59: 346-53
MultiD_Analyses_AB. 2019.
Munroe DJ, Harris TJR. 2010. Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology* 28: 426-28
Muto N, Matsumoto T. 2022. CRISPR/Cas9-mediated genome editing of RsGL1a and RsGL1b in radish (Raphanus sativus L.). *Front Plant Sci* 13: 951660
Nayidu NK, Kagale S, Taheri A, Withana-Gamage TS, Parkin IA, et al. 2014. Comparison of five major trichome regulatory genes in Brassica villosa with orthologues within the Brassicaceae. *PloS one* 9: e95877
Negrón-Ortiz V. 2007. Chromosome numbers, nuclear DNA content, and polyploidy in Consolea (Cactaceae), an endemic cactus of the Caribbean Islands. *American journal of botany* 94: 1360-70
Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, USA.
Nekrutenko A, Makova KD, Li WH. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* 12: 198-202
Nikolov LA, Shushkov P, Nevado B, Gan X, Al-Shehbaz IA, et al. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist* 222: 1638-51
Noda K-i, Glover BJ, Linstead P, Martin C. 1994. Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor. *Nature* 369: 661-64
Nougué O, Corbi J, Ball SG, Manicacci D, Tenaillon MI. 2014. Molecular evolution accompanying functional divergence of duplicated genes along the plant starch biosynthesis pathway. *BMC Evol Biol* 14: 1-16
Nuruzzaman M, Gupta M, Zhang C, Wang L, Xie W, et al. 2008. Sequence and expression analysis of the thioredoxin protein gene family in rice. *Molecular Genetics and Genomics* 280: 139
Ogata K, Hojo H, Aimoto S, Nakai T, Nakamura H, et al. 1992. Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proceedings of the National Academy of Sciences* 89: 6428-32
Ohno S. 1970. *Evolution by Gene Duplication*. Springer.
Ohno S. 2013. *Evolution by gene duplication*. Springer Science & Business Media.

Okamoto S, Negishi K, Toyama Y, Ushijima T, Morohashi K. 2020. Leaf Trichome Distribution Pattern in Arabidopsis Reveals Gene Expression Variation Associated with Environmental Adaptation. *Plants* 9: 909

Oppenheimer DG, Herman PL, Sivakumaran S, Esch J, Marks MD. 1991. A myb gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules. *Cell* 67: 483-93

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, et al. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in genetics* 19: 141-47

Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, et al. 2008. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res* 36: W163-9

Ownbey M, McCollum GD. 1953. CYTOPLASMIC INHERITANCE AND RECIPROCAL AMPHIPLOIDY IN TRAGOPOGON. *American journal of botany* 40: 788-96

Paffendorf BA, Qassrawi R, Meys AM, Trimborn L, Schrader A. 2020. TRANSPARENT TESTA GLABRA 1 participates in flowering time regulation in Arabidopsis thaliana. *PeerJ* 8: e8303

Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of Gene Duplication in Plants. *Plant Physiology* 171: 2294-316

Paterson A, Wang X, Tang H, Lee T. 2012. *Synteny and genomic rearrangements*. Springer.

Patra B, Pattanaik S, Yuan L. 2013. Ubiquitin protein ligase 3 mediates the proteasomal degradation of GLABROUS 3 and ENHANCER OF GLABROUS 3, regulators of trichome development and flavonoid biosynthesis in Arabidopsis. *The Plant Journal* 74: 435-47

Pattanaik S, Patra B, Singh SK, Yuan L. 2014. An overview of the gene regulatory network controlling trichome development in the model plant, Arabidopsis. *Front Plant Sci* 5: 259

Payne CT, Zhang F, Lloyd AM. 2000. GL3 Encodes a bHLH Protein That Regulates Trichome Development in Arabidopsis Through Interaction With GL1 and TTG1. *Genetics* 156: 1349-62

Payne T, Clement J, Arnold D, Lloyd A. 1999. Heterologous myb genes distinct from GL1 enhance trichome production when overexpressed in Nicotiana tabacum. *Development* 126: 671-82

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome Size Diversity and Its Impact on the Evolution of Land Plants. *Genes* 9: 88

Pérez-Pérez JM, Candela H, Micol JL. 2009. Understanding synergy in genetic interactions. *Trends Genet* 25: 368-76

Pesch M, Dartan B, Birkenbihl R, Somssich IE, Hülskamp M. 2014. Arabidopsis TTG2 regulates TRY expression through enhancement of activator complex-triggered activation. *Plant Cell* 26: 4067-83

Pesch M, Hülskamp M. 2009. One, two, three…models for trichome patterning in Arabidopsis? *Current Opinion in Plant Biology* 12: 587-92

Pesch M, Schultheiß I, Klopffleisch K, Uhrig JF, Koegl M, et al. 2015. TRANSPARENT TESTA GLABRA1 and GLABRA1 Compete for Binding to GLABRA3 in Arabidopsis. *Plant Physiology* 168: 584-97

Pfaffl MW. 2012. Quantification strategies in real-time polymerase chain reaction. *Quantitative real-time PCR Appl Microbiol*: 53-62

Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology letters* 26: 509-15

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, et al. 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* 345: 1250091

Pickett FB, Meeks-Wagner DR. 1995. Seeing double: appreciating genetic redundancy. *The Plant Cell* 7: 1347-56

Pires JC, Zhao J, Schranz ME, Leon EJ, Quijada PA, et al. 2004. Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). *Biological Journal of the Linnean Society* 82: 675-88

Pireyre M, Burow M. 2015. Regulation of MYB and bHLH Transcription Factors: A Glance at the Protein Level. *Molecular Plant* 8: 378-88
Pophaly SD, Tellier A. 2015. Population level purifying selection and gene expression shape subgenome evolution in maize. *Molecular biology and evolution* 32: 3226-35
Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, et al. 2015. Codon optimality is a major determinant of mRNA stability. *Cell* 160: 1111-24
Price JP, Wagner WL. 2004. Speciation in Hawaiian Angiosperm Lineages: Cause, Consequence, and Mode. *Evolution* 58: 2185-200
Price RA, Al-Shehbaz IA, O'Kane SL. 2001. Beringia (Brassicaceae), a New Genus of Arabidopsoid Affinities from Russia and North America. *Novon* 11: 332-36
Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. *Nature Biotechnology* 27: 847
Qi B, Huang W, Zhu B, Zhong X, Guo J, et al. 2012. Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. *BMC Biology* 10: 3
Qian W, Ma D, Xiao C, Wang Z, Zhang J. 2012. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell reports* 2: 1399-410
Qiao X, Li Q, Yin H, Qi K, Li L, et al. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome biology* 20: 1-23
Ramachandran GN, Ramakrishnan C, Sasisekharan V. 1963. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology* 7: 95-9
Ramsay NA, Glover BJ. 2005. MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. *Trends in Plant Science* 10: 63-70
Ramsey J, Ramsey TS. 2014. Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130352
Ramsey J, Schemske DW. 1998. PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. *Annual Review of Ecology and Systematics* 29: 467-501
Ramsey J, Schemske DW. 2002. Neopolyploidy in Flowering Plants. *Annual Review of Ecology and Systematics* 33: 589-639
Rapp RA, Udall JA, Wendel JF. 2009. Genomic expression dominance in allopolyploids. *BMC Biology* 7: 18
Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol* 5: 28
Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol* 16: 266-74
Ray A, Lindahl E, Wallner B. 2012. Improved model quality assessment using ProQ2. *BMC Bioinformatics* 13: 224
Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, et al. 2014. Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. *Genome biology and evolution* 6: 559-71
Rerie WG, Feldmann KA, Marks MD. 1994. The GLABRA2 gene encodes a homeo domain protein required for normal trichome development in Arabidopsis. *Genes & development* 8: 1388-99
Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, et al. 2000. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *science* 290: 2105-10
Rinehart JA, Petersen MW, John ME. 1996. Tissue-specific and developmental regulation of cotton gene FbL2A (demonstration of promoter activity in transgenic plants). *Plant Physiology* 112: 1331-41
Roche. 2001. Roche Applied Science Technical Note No. LC 13/2001
Rodriguez MA, Vermaak D, Bayes JJ, Malik HS. 2007. Species-specific positive selection of the male-specific lethal complex that participates in dosage compensation in Drosophila. *Proceedings of the National Academy of Sciences* 104: 15412-17

Román-Palacios C, Medina CA, Zhan SH, Barker MS. 2021. Animal chromosome counts reveal a similar range of chromosome numbers but with less polyploidy in animals compared to flowering plants. *Journal of Evolutionary Biology* 34: 1333-39

Romero I, Fuertes A, Benito M, Malpica J, Leyva A, Paz-Ares J. 1998. More than 80R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. *The Plant journal: for cell and molecular biology* 14: 273-84

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, et al. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 308: 58-73

Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, et al. 2013. The fate of duplicated genes in a polyploid plant genome. *The Plant Journal* 73: 143-53

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, et al. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution* 34: 3299-302

Samuel Yang S, Cheung F, Lee JJ, Ha M, Wei NE, et al. 2006. Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *The Plant Journal* 47: 761-75

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci. USA* 74: 5463-67

Sankoff D, Zheng C. 2018. Whole genome duplication in plants: implications for evolutionary analysis. *Comparative Genomics: Methods and Protocols*: 291-315

Satina S, Blakeslee AF. 1937. Chromosome behavior in triploids of Datura stramonium. I. The male gametophyte. *American journal of botany*: 518-27

Sato Y, Kudoh H. 2015. Tests of associational defence provided by hairy plants for glabrous plants of Arabidopsis halleri subsp. gemmifera against insect herbivores. *Ecological Entomology* 40: 269-79

Sattler MC, Carvalho CR, Clarindo WR. 2016. The polyploidy and its key role in plant breeding. *Planta* 243: 281-96

Schellmann S, Schnittger A, Kirik V, Wada T, Okada K, et al. 2002. TRIPTYCHON and CAPRICE mediate lateral inhibition during trichome and root hair patterning in Arabidopsis. *Embo j* 21: 5036-46

Schenk JJ. 2021. The Next Generation of Adaptive Radiation Studies in Plants. *International Journal of Plant Sciences* 182: 245-62

Schiefelbein J, Zheng X, Huang L. 2014. Regulation of epidermal cell fate in Arabidopsis roots: the importance of multiple feedback loops. *Front Plant Sci* 5

Schiessl S. 2020. Regulation and Subfunctionalization of Flowering Time Genes in the Allotetraploid Oil Crop Brassica napus. *Front Plant Sci* 11

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* 43: e37-e37

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, et al. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47: 868-76

Schmidtke P, Le Guilloux V, Maupetit J, Tuffi¿ ½ry P. 2010. Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic acids research* 38: W582-W89

Schnable JC, Freeling M. 2011. Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PloS one* 6: e17855

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences* 108: 4069-74

Schnittger A, Folkers U, Schwab B, Jürgens G, Hülskamp M. 1999. Generation of a Spacing Pattern: The Role of TRIPTYCHON in Trichome Patterning in Arabidopsis. *The Plant Cell* 11: 1105-16

Schranz ME, Mitchell-Olds T. 2006. Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae. *The Plant Cell* 18: 1152-65

Schwab B, Folkers U, Ilgenfritz H, Hülskamp M. 2000. Trichome morphogenesis in Arabidopsis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 355: 879-83
Segraves KA. 2017. The effects of genome duplications in a community context. *New Phytologist* 215: 57-69
Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA. 2015. Cytonuclear coordination is not immediate upon allopolyploid formation in Tragopogon miscellus (Asteraceae) allopolyploids. *PloS one* 10: e0144339
Sémon M, Wolfe KH. 2007. Consequences of genome duplication. *Current opinion in genetics & development* 17: 505-12
Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, et al. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* 20: 633-43
Serna L, Martin C. 2006. Trichomes: different regulatory networks lead to convergent structures. *Trends in plant science* 11: 274-80
Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309: 1728-32
Shi X, Zhang C, Ko DK, Chen ZJ. 2015. Genome-Wide Dosage-Dependent and -Independent Regulation Contributes to Gene Expression and Evolutionary Novelty in Plant Polyploids. *Molecular Biology and Evolution* 32: 2351-66
Shimeld SM. 1999. Gene function, gene networks and the fate of duplicated genes. *Seminars in Cell & Developmental Biology* 10: 549-53
Siena LA, Ortiz JPA, Calderini O, Paolocci F, Cáceres ME, et al. 2016. An apomixis-linked ORC3-like pseudogene is associated with silencing of its functional homolog in apomictic Paspalum simplex. *Journal of Experimental Botany* 67: 1965-78
Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends in Genetics* 34: 532-44
Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 99: 13627-32
Singh ND, Kumar S, Daniell H. 2016. Expression of β-glucosidase increases trichome density and artemisinin content in transgenic Artemisia annua plants. *Plant biotechnology journal* 14: 1034-45
Soltis DE, Buggs RJ, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* 59: 1387-403
Soltis DE, Soltis PS. 1999. Polyploidy: recurrent formation and genome evolution. *Trends in ecology & evolution* 14: 348-52
Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant speciation. *New Phytologist* 161: 173-91
Soltis DE, Soltis PS, Tate JA. 2004. Advances in the study of polyploidy since Plant speciation. *New Phytologist* 161: 173-91
Soltis DE, Visger CJ, Soltis PS. 2014. The polyploidy revolution then...and now: Stebbins revisited. *American journal of botany* 101: 1057-78
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current opinion in genetics & development* 35: 119-25
Soltis PS, Soltis DE. 2000. The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences* 97: 7051-57
Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Current opinion in plant biology* 30: 159-65
Song K, Lu P, Tang K, Osborn TC. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proceedings of the National Academy of Sciences* 92: 7719-23
Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P. 1990. Conformations of disulfide bridges in proteins. *International Journal of Peptide and Protein Research* 36: 147-55
Stebbins Jr CL. 1950. Variation and evolution in plants. *Variation and evolution in plants.*

Stracke R, Werber M, Weisshaar B. 2001. The R2R3-MYB gene family in Arabidopsis thaliana. *Current Opinion in Plant Biology* 4: 447-56
Stuessy TF, Crawford DJ. 1998. Chromosomal stasis during speciation in angiosperms of oceanic islands In *Evolution and Speciation of Island Plants*, ed. M Ono, TF Stuessy, pp. 307-24. Cambridge: Cambridge University Press
Sumitomo K, Shirasawa K, Isobe S, Hirakawa H, Hisamatsu T, et al. 2019. Genome-wide association study overcomes the genome complexity in autohexaploid chrysanthemum and tags SNP markers onto the flower color genes. *Scientific Reports* 9: 13947
Suo J, Liang X, Pu L, Zhang Y, Xue Y. 2003. Identification of GhMYB109 encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (Gossypium hirsutum L.). *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* 1630: 25-34
Symonds VV, Hatlestad G, Lloyd AM. 2011. Natural allelic variation defines a role for ATMYC1: trichome cell fate determination. *PLoS Genet* 7: e1002069
Szymanski DB, Jilk RA, Pollock SM, Marks MD. 1998. Control of GL2 expression in Arabidopsis leaves and trichomes. *Development* 125: 1161-71
Szymanski DB, Lloyd AM, Marks MD. 2000. Progress in the molecular genetic analysis of trichome initiation and morphogenesis in Arabidopsis. *Trends Plant Sci* 5: 214-9
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-95
Tamayo-Ordóñez M, Espinosa-Barrera L, Tamayo-Ordóñez Y, Ayil-Gutiérrez B, Sánchez-Teyer L. 2016. Advances and perspectives in the generation of polyploid plant species. *Euphytica* 209: 1-22
Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38: 3022-27
Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, et al. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics* 190: 1563-74
Tate JA, Soltis DE, Soltis PS. 2005. Polyploidy in plants In *The evolution of the genome*, pp. 371-426: Elsevier
Telfer A, Bollman KM, Poethig RS. 1997. Phase change and the regulation of trichome distribution in Arabidopsis thaliana. *Development* 124: 645-54
Teo ZL, Savas P, Loi S. 2016. Gene Expression Analysis: Current Methods In *Molecular Pathology in Cancer Research*, ed. SR Lakhani, SB Fox, pp. 107-36. New York, NY: Springer New York
Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 16: 934-46
Thornton B, Basu C. 2011. Real-time PCR (qPCR) primer design using free online software. *Biochem Mol Biol Educ* 39: 145-54
Throude M, Bolot S, Bosio M, Pont C, Sarda X, et al. 2009. Structure and expression analysis of rice paleo duplications. *Nucleic acids research* 37: 1248-59
Toledo-Ortiz G, Huq E, Quail PH. 2003. The Arabidopsis basic/helix-loop-helix transcription factor family. *The Plant Cell* 15: 1749-70
Tominaga-Wada R, Ishida T, Wada T. 2011. New insights into the mechanism of development of Arabidopsis root hairs and trichomes. *International review of cell and molecular biology* 286: 67-106
Tsuji J, Coe L. 2013. The glabra1 mutation affects the stomatal patterning of Arabidopsis thaliana rosette leaves. *bios* 84: 92-97
Tyagi S, Sri T, Singh A, Mayee P, Shivaraj SM, et al. 2019. SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 influences flowering time, lateral branching, oil quality, and seed yield in Brassica juncea cv. Varuna. *Functional & Integrative Genomics* 19: 43-60

Udall JA, Quijada PA, Osborn TC. 2005. Detection of chromosomal rearrangements derived from homeologous recombination in four mapping populations of Brassica napus L. *Genetics* 169: 967-79
Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, et al. 2012. Primer3--new capabilities and interfaces. *Nucleic acids research* 40: e115-e15
Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* 35: W71-W74
Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* 33: 11-26
Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009a. The flowering world: a tale of duplications. *Trends in Plant Science* 14: 680-88
Van de Peer Y, Maere S, Meyer A. 2009b. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10: 725-32
Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* 18: 411-24
van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* 30: 418-26
VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, et al. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Molecular systems biology* 6: 429
Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, et al. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology* 3: 1-12
VanGuilder HD, Vrana KE, Freeman WM. 2008. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* 44: 619-26
Vanin EF. 1985. Processed pseudogenes: characteristics and evolution. *Annual review of genetics* 19: 253-72
Veitia RA, Bottani S, Birchler JA. 2013. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends in Genetics* 29: 385-93
Videvall E, Sletvold N, Hagenblad J, Ågren J, Hansson B. 2015. Strong Maternal Effects on Gene Expression in Arabidopsis lyrata Hybrids. *Molecular Biology and Evolution* 33: 984-94
Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. *Science* 290: 2114-17
Vries Hd. 1901. Die Mutationstheorie: Versuche und Beobachtungen über die Entstehung von Arten im Pflanzenreich. *(No Title)*
Wada T, Tachibana T, Shimura Y, Okada K. 1997. Epidermal cell differentiation in Arabidopsis determined by a Myb homolog, CPC. *Science* 277: 1113-6
Walker A, Marks M. 2000. Trichome initiation in Arabidopsis.
Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N, et al. 1999. The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. *The Plant cell* 11: 1337-50
Wang J, Roe B, Macmil S, Yu Q, Murray JE, et al. 2010. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11: 261
Wang J, Tian L, Lee H-S, Wei NE, Jiang H, et al. 2006. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics* 172: 507-17
Wang J, Tian L, Madlung A, Lee HS, Chen M, et al. 2004a. Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics* 167: 1961-73
Wang L, Si W, Yao Y, Tian D, Araki H, Yang S. 2012. Genome-wide survey of pseudogenes in 80 fully re-sequenced Arabidopsis thaliana accessions. *PloS one* 7: e51769
Wang S, Adams KL. 2015. Duplicate gene divergence by changes in microRNA binding sites in Arabidopsis and Brassica. *Genome biology and evolution* 7: 646-55

Wang S, Chen J-G. 2008. Arabidopsis Transient Expression Analysis Reveals that Activation of GLABRA2 May Require Concurrent Binding of GLABRA1 and GLABRA3 to the Promoter of GLABRA2. *Plant and Cell Physiology* 49: 1792-804
Wang S, Chen J-G. 2014. Regulation of cell fate determination by single-repeat R3 MYB transcription factors in Arabidopsis. *Front Plant Sci* 5
Wang S, Wang J-W, Yu N, Li C-H, Luo B, et al. 2004b. Control of plant trichome development by a cotton fiber MYB gene. *The Plant Cell* 16: 2323-34
Wang X, Dong Q, Li X, Yuliang A, Yu Y, et al. 2017. Cytonuclear variation of Rubisco in synthesized rice hybrids and allotetraploids. *The Plant Genome* 10: plantgenome2017.05.0041
Wang X, Morton JA, Pellicer J, Leitch IJ, Leitch AR. 2021a. Genome downsizing after polyploidy: mechanisms, rates and selection pressures. *The Plant Journal* 107: 1003-15
Wang X, Shen C, Meng P, Tan G, Lv L. 2021b. Analysis and review of trichomes in plants. *BMC plant biology* 21: 1-11
Wang X, Wang H, Wang J, Sun R, Wu J, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* 43: 1035-39
Wang X, Zhang N, Zhao Y, Wang J. 2021c. A New Method for Recognizing Protein Complexes Based on Protein Interaction Networks and GO Terms. *Front Genet* 12: 792265
Wang Y, Ficklin SP, Wang X, Feltus FA, Paterson AH. 2016. Large-scale gene relocations following an ancient genome triplication associated with the diversification of core eudicots. *PloS one* 11: e0155637
Wang Y, Zhou Q, Meng Z, Abid MA, Wang Y, et al. 2022a. Multi-Dimensional Molecular Regulation of Trichome Development in Arabidopsis and Cotton. *Front Plant Sci* 13: 892381
Wang Z, Yang J, Cheng F, Li P, Xin X, et al. 2022b. Subgenome dominance and its evolutionary implications in crop domestication and breeding. *Horticulture Research* 9: uhac090
Wang Z, Yang Z, Li F. 2019. Updates on molecular mechanisms in the development of branched trichome in Arabidopsis and nonbranched in cotton. *Plant Biotechnology Journal* 17: 1706-22
Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20: 2138-39
Wei Z, Cheng Y, Zhou C, Li D, Gao X, et al. 2019. Genome-wide identification of direct targets of the TTG1–bHLH–MYB complex in regulating trichome formation and flavonoid accumulation in Arabidopsis thaliana. *International Journal of Molecular Sciences* 20: 5014
Weiss-Schneeweiss H, Emadzade K, Jang T-S, Schneeweiss GM. 2013. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenetic and genome research* 140: 137-50
Wendel JF. 2000. Genome evolution in polyploids. *Plant molecular evolution*: 225-49
Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr Opin Genet Dev* 49: 1-7
Winkler H. 1916. Uber die experimentelle Erzeugung von Pflanzen mit abweichen den Chromosomenzahlen. Zeitschr. f. *Bot. Bd* 8
Wolf JB. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol Ecol Resour* 13: 559-72
Wu X, Qi X. 2010. Genes encoding hub and bottleneck enzymes of the Arabidopsismetabolic network preferentially retain homeologs through whole genome duplication. *BMC Evol Biol* 10: 145
Wu Y, Sun Y, Wang X, Lin X, Sun S, et al. 2016. Transcriptome shock in an interspecific F1 triploid hybrid of Oryza revealed by RNA sequencing. *Journal of Integrative Plant Biology* 58: 150-64

Xiao C, Zhang H, Xie F, Pan Z-Y, Qiu W-M, et al. 2022. Evolution, gene expression, and protein–protein interaction analyses identify candidate CBL-CIPK signalling networks implicated in stress responses to cold and bacterial infection in citrus. *BMC Plant Biology* 22: 1-17

Xiong Z, Gaeta RT, Pires JC. 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. *Proceedings of the National Academy of Sciences* 108: 7908-13

Xu W, Dubos C, Lepiniec L. 2015. Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. *Trends Plant Sci* 20: 176-85

Xuan L, Yan T, Lu L, Zhao X, Wu D, et al. 2020. Genome-wide association study reveals new genes involved in leaf trichome formation in polyploid oilseed rape (Brassica napus L.). *Plant, Cell & Environment* 43: 675-91

Yan M, Liu X, Guan C, Liu L, Xiang J, et al. 2014. Cloning of TTG1 gene and PCR identification of genomes A, B and C in Brassica species. *Genetica* 142: 169-76

Yang C, Ye Z. 2013. Trichomes as models for studying plant cell differentiation. *Cellular and molecular life sciences* 70: 1937-48

Yang J, Liu D, Wang X, Ji C, Cheng F, et al. 2016. The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature genetics* 48: 1225-32

Yates CM, Filippis I, Kelley LA, Sternberg MJE. 2014. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *Journal of molecular biology* 426: 2692-701

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13: 134

Yim WC, Lee B-M, Jang CS. 2009. Expression diversity and evolutionary dynamics of rice duplicate genes. *Molecular Genetics and Genomics* 281: 495-95

Yogeeswaran K, Voelckel C, Joly S, Heenan PB. 2011. Pachycladon. 227-49

Yoo MJ, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110: 171-80

Yu Z, Zheng C, Albert VA, Sankoff D. 2020. Excision dominates pseudogenization during fractionation after whole genome duplication and in gene loss after speciation in plants. *Frontiers in Genetics* 11: 603056

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18: 821-29

Zhang B, Chopra D, Schrader A, Hülskamp M. 2019a. Evolutionary comparison of competitive protein-complex formation of MYB, bHLH, and WDR proteins in plants. *Journal of Experimental Botany* 70: 3197-209

Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. 2003. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. *Development* 130: 4859-69

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18: 292-98

Zhang J. 2012. Genetic redundancies and their evolutionary maintenance. *Evolutionary systems biology*: 279-300

Zhang L, He J, He H, Wu J, Li M. 2021. Genome-wide unbalanced expression bias and expression level dominance toward Brassica oleracea in artificially synthesized intergeneric hybrids of Raphanobrassica. *Horticulture Research* 8

Zhang M, Liu X-K, Fan W, Yan D-F, Zhong N-S, et al. 2018. Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from Camellia. *Genome* 61: 477-85

Zhang Y, Zhu H, Shao C, Cai F, Zhang J, Bao M. 2019b. PaMYB82 from Platanus acerifolia regulates trichome development in transgenic Arabidopsis. *Plant Science* 287: 110177

Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. 2006. KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genomics, Proteomics & Bioinformatics* 4: 259-63

Zhao B, Liu L, Tan D, Wang J. 2010. Analysis of phylogenetic relationships of Brassicaceae species based on Chs sequences. *Biochemical Systematics and Ecology* 38: 731-39

Zhao H, Li X, Ma L. 2012a. Basic helix-loop-helix transcription factors and epidermal cell fate determination in Arabidopsis. *Plant Signal Behav* 7: 1556-60

Zhao H, Wang X, Zhu D, Cui S, Li X, et al. 2012b. A Single Amino Acid Substitution in IIIf Subfamily of Basic Helix-Loop-Helix Transcription Factor AtMYC1 Leads to Trichome and Root Hair Patterning Defects by Abolishing Its Interaction with Partner Proteins in Arabidopsis. *Journal of Biological Chemistry* 287: 14109-21

Zhao L, Gao L, Wang H, Chen X, Wang Y, et al. 2013. The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis. *Funct Integr Genomics* 13: 75-98

Zhao M, Morohashi K, Hatlestad G, Grotewold E, Lloyd A. 2008. The TTG1-bHLH-MYB complex controls trichome cell fate and patterning through direct targeting of regulatory loci. *Development* 135: 1991-99

Zheng D, Ye W, Song Q, Han F, Zhang T, Chen ZJ. 2016a. Histone modifications define expression bias of homoeologous genomes in allotetraploid cotton. *Plant physiology* 172: 1760-71

Zheng K, Tian H, Hu Q, Guo H, Yang L, et al. 2016b. Ectopic expression of R3 MYB transcription factor gene OsTCL1 in Arabidopsis, but not rice, affects trichome and root hair formation. *Scientific reports* 6: 19254

Zhou L, Zheng K, Wang X, Tian H, Wang X, Wang S. 2014. Control of trichome formation in Arabidopsis by poplar single-repeat R3 MYB transcription factors. *Front Plant Sci* 5: 262

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF. 2004. Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. *Plant J* 40: 22-34

Zou C, Lehti-Shiu MD, Thibaud-Nissen Fo, Prakash T, Buell CR, Shiu S-H. 2009. Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice *Plant Physiology* 151: 3-15