A transcriptomics approach to understanding polymorphic and transcript level differences linked to isoquinoline alkaloid production in triploid varieties of *Narcissus pseudonarcissus*

Jane Pulman

The University of Liverpool

September 2014

UNIVERSITY OF LIVERPOOL

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by Jane Pulman

Acknowledgements

I would like to make the following acknowledgements;

I would first and foremost like to thank my supervisors, Dr M Jones for her unwavering support and critical evaluation of my thesis and Prof. A Hall for helping me on the road to bioinformatics discovery.

I would like to thank Travis Banks for the opportunity to study with him, without his patience and tutorage I would not have been able to complete my thesis. I am eternally grateful for not only the skill he has taught me but also his guidance and evaluation of my thesis.

Vineland Research and Innovation Centre in particular Dr D Somers and Dr D Liscombe for not only the opportunity to spend time working in such a wonderful research centre but also for their continued support and guidance.

Dr X Chang and Prof. T Walker for their support, samples, data and guidance.

The OSVH Trust for the PhD sponsorship and funding.

A special thank you to Dr C Lofthouse and (soon to be) Dr Jade Waller without your support, guidance and coffee breaks I don't think I would of made it.

To everyone in lab G, past and present for putting up with a chemist trying to make it in a biology lab, especially Jean Wood for putting up with my mess.

To friends and family, thank you for your continued belief in me and putting up with my thesis induced mood swings. In particular I'd like to thank my Mam, Dad, Dave and Kat for your support and willingness to tell me to get a grip and just do it. You always said I'd be a Doctor Mam and no Dad I still don't work on dandelions.

Grandma (both of you), thank you for your kind words of encouragement, and Grandma Ella thanks for making me laugh and the never-ending supply of marshmallows.

Jonathan thanks for coming on all my "little" adventures with me, can't wait for the next one.

Beatrice, Katy, Louise, Jen, Laura, Nick, Gwen, Phil, Jim, Justin, Jules, Jo, Rosie, Ellen, Kathy, team EVO-fit (tough mudder) and anyone I've forgotten thanks for making my time in Vineland and Liverpool amazing.

I would finally like to thank Rob for looking after me and making me nice teas while I was writing up. Your support has really got me through.

Emma and Alex last but not least thanks for all the texts, postcards, songs, and good times that made my stresses just disappear.

This Thesis is dedicated in loving memory to Jack Auton, my biggest supporter.

Table of contents

LIST OF FIGURES IX

LIST OF TABLES XI

ABSTRACT XII

Abbreviations xiii

1 CHAPTER ONE: THE USE OF SECOND-GENERATION SEQUENCING TECHNOLOGIES FOR THE STUDY OF MEDICINALLY IMPORTANT ALKALOIDS - A CASE STUDY IN GALANTHAMINE PRODUCTION IN DAFFODILS 1

1.1 Introduction 1

1.2 Daffodil- Narcissus pseudonarcissus 4

1.3 Amaryllidaceae alkaloids – History of their medicinal properties 5
1.3.1 Galanthamine – as a folk medicine 7
1.3.2 Other Amaryllidaceae alkaloids with potential medical uses: Narciclasine and Lycorine 7

1.4 Galanthamine 8
1.4.1 As a proven drug in Alzheimer's disease 8
1.4.2 Chemical synthesis 10
1.4.3 Biosynthesis 11

1.5 The use of functional genomics in plant science 13
1.5.1 The development of Second-generation sequencing technologies 13
1.5.1.1 454 pyrosequencing 14
1.5.1.2 Illumina sequencing 15
1.5.1.3 Ion Torrent 15
1.5.2 From genome to transcriptome 16
1.5.3 The use of second-generation sequencing in plant science 17
1.5.4 Transcriptomic studies on non-model plants 18
1.5.5 The future of sequencing technologies 19
1.5.6 Transcriptomics in the study of the biosynthesis of plant alkaloids and other high value compounds 20
1.5.7 Transcript expression analysis in the search for agronomically and medicinally important compounds 22
1.5.8 SNP analysis in the search for high value compounds 23

1.6 Conclusion and Hypothesis 25
1.6.1 Conclusion 25
1.6.2 Hypothesis 25

2 CHAPTER TWO - CONSTRUCTION OF A REFERENCE TRANSCRIPTOME 26

2.1 Introduction 26
2.1.1 The use of Second-generation sequencing for de-novo assembly of transcriptomes in plants that produce medicinally important metabolites 26
2.1.2 454 Sequencing technology 26
2.1.3 Assemblers for *de novo* assembly of second-generation transcriptome data 26
2.1.4 Overlap assembly methods 27

2.1.5 Analysing transcriptome *de novo* assemblies 27
2.1.6 Annotation of *de novo* assemblies – The use of BLAST searches against known databases 28
2.1.7 RNA extraction from plants with high phenol, sugar and secondary metabolite levels 29
2.1.8 cDNA library preparation 29

2.2 Aims and Objectives 31
2.2.1 Overview of aims and objectives 31
2.2.2 Data assembly and annotation 31

2.3 Methods 32
2.3.1 Plant material for reference transcriptome 32
2.3.2 Basal plate extraction 32
2.3.3 Production of cDNA library 32
2.3.3.1 Total RNA extraction – CTAB method 32
2.3.3.2 mRNA isolation 33
2.3.3.3 rRNA depletion 33
2.3.3.4 mRNA selection 33
2.3.3.5 cDNA Library preparation 34
2.3.4 Creation of a reference transcriptome 34
2.3.4.1 454 Pyrosequencing 34
2.3.4.2 Assembly of reads from sequencing of barcoded libraries 34
2.3.4.3 Re-sequencing of sample 2 and full assembly from all GS FLX data 34
2.3.4.4 Manual annotation and the creation of an automated annotation pipeline for the joint and full assemblies using “full_annotation.pl” 35

2.4 Results 37
2.4.1 Creation of a cDNA library 37
2.4.1.1 RNA extraction – CTAB Method 37
2.4.1.2 mRNA isolation 38
2.4.1.2.1 rRNA depletion of sample 1 38
2.4.1.2.2 mRNA selection of sample 2 39
2.4.1.3 cDNA library preparation 40
2.4.1.3.1 Sample 1 40
2.4.1.3.2 Sample 2 40
2.4.2 Production of a reference transcriptome 41
2.4.2.1 Assembly of reads from sequencing of the barcoded libraries 41
2.4.2.2 Mapping of barcoded libraries to the joint assembly 41
2.4.2.3 Re-sequencing of the sample 2 43
2.4.2.4 Assembly of the re-sequenced data with the initial barcoded data (full assembly) 43
2.4.2.5 Mapping of thee sequencing samples to full assembly 43
2.4.3 Annotation and creation of an automated annotation pipeline for the joint and full assemblies 44
2.4.3.1 Joint assembly 44
2.4.3.2 Full assembly 45

2.5 Discussion 46

3 CHAPTER THREE: IMPLEMENTING SECOND-GENERATION SEQUENCING DATA FOR DISCOVERING POLYMORPHIC AND TRANSCRIPT LEVEL DIFFERENCES BETWEEN TWO VARIETIES OF *NARCISSUS PSEUDONARCISSUS* WITH DISTINCT DIFFERENCES IN GALANTHAMINE LEVELS. 49

3.1 Introduction 49
3.1.1 Illumina Sequencing 49
3.1.2 Assembly of Illumina reads 50
3.1.2.1 Assemblers that use de Bruijn graphs 50
3.1.2.2 Comparison of assemblers 51

3.1.2.3 Trinity 52
3.1.2.4 SOAPdenovo-Trans 53
3.1.3 Aligning short reads back to reference transcriptomes – comparison of available tools 54
3.1.4 Hash Tables 56
3.1.5 Burrows Wheeler Transformation (BWT) aligners 56
3.1.6 Comparison of available aligners 57
3.1.6.1 Bowtie or BWA 58
3.1.7 Analysis of polymorphic differences 58
3.1.7.1 Identification of SNPs in polyploids 59
3.1.7.2 VarScan 61
3.1.7.3 Pileup_parser.pl perl script for determining all possible allele variation 62
3.1.7.4 The use of pileup_parser.pl to predict ploidy levels at the nucleotide level 64
3.1.8 Analysis of transcript level differences between individuals from two varieties of *Narcissus pseudonarcissus* 65
3.1.8.1 BitSeq 65
3.1.9 Transposon element variation within plants 66
3.1.9.1 TransposonPSI 67
3.1.9.2 TREP 67
3.1.9.3 BMTagger 68

3.2 Aims and Objectives 69

3.3 Methodology 70
3.3.1 Illumina library preparation, sequencing, assembly and annotation 70
3.3.1.1 Extraction of RNA from N. pseudonarcissus var Andrew's Choice 70
3.3.1.2 Illumina cDNA library preparation for N. pseudonarcissus var. Carlton and N. pseudonarcissus var. Andrew's Choice 71
3.3.1.3 Analysis of Sequence data 73
3.3.1.4 Basic script and program running 74
3.3.1.5 Trimming and Filtering of raw reads 74
3.3.1.6 Comparison of de novo assembly using Trinity and SOAPdenovo-Trans 75
3.3.1.7 Mapping of trimmed Illumina reads to the Trinity assembly, SOAPdenovo-Trans and the 454 references 75
3.3.2 SNP calling using custom SNP caller for polyploid samples and the prediction of ploidy level 76
3.3.3 Transcript level differences determined using BitSeq 77
3.3.4 Annotation of Trinity assembly 77
3.3.5 Removal of transposon elements from raw reads – comparison of three methods 78
3.3.5.1 BMTagger 78
3.3.5.2 TransposonPSI 78
3.3.5.3 BLAST against TREP 78
3.3.6 Re-assembly of reads post TE removal 78
3.3.7 Annotation of the re-assembled data 79
3.3.8 SNP and transcript level differences re-evaluated post transposon removal 79

3.4 Results 80
3.4.1 Extraction of RNA 80
3.4.1.1 Andrew's Choice 80
3.4.1.2 Carlton 80
3.4.2 RNA-Seq library preparation 80
3.4.2.1 Total RNA analysis 80
3.4.2.2 cDNA library preparation 83
3.4.2.3 Trimming and Filtering of raw reads 85
3.4.2.4 Comparison of de novo assembly using Trinity and SOAPdenovo-Trans 89
3.4.2.5 Mapping of trimmed Illumina reads to the Trinity, SOAPdenovo-Trans and 454 assemblies 89
3.4.2.6 SNP calling using custom SNP caller and VarScan 90
3.4.2.7 The prediction of ploidy level in Carlton and Andrew's Choice 91
3.4.2.8 Transcript level differences determined using BitSeq 91
3.4.2.9 Annotation of Trinity assembly 92

3.4.2.10 Removal of transposon elements from raw reads – comparison of three methods 92
3.4.2.11 Assembly comparison of the three TE removal methods 94
3.4.2.12 Mapping of TE removed reads to both the 454 reference and the TE removed Trinity assembly 95
3.4.2.13 SNP calling in both the Original and Post-TE removal assembly 96
3.4.2.14 Transcript level differences 96
3.4.2.15 Annotation post TE removal 97

3.5 Discussion 98
3.5.1 Production of cDNA libraries and read quality control 98
3.5.2 Assembly of Illumina reads 99
3.5.3 Removal of TEs 101
3.5.4 Annotation of Transcripts 102
3.5.5 SNP and transcript level difference analysis 103
3.5.6 Bioinformatics approach to predicting ploidy level of Carlton and Andrew's Choice 105

4 CHAPTER FOUR: ANALYSIS OF TRANSCRIPTOME DATA TO DETERMINE PUTATIVE TRANSCRIPTS LINKED TO AMARYLLIDACEAE ALKALOID BIOSYNTHESIS 106

4.1 Introduction 106
4.1.1 Methods of annotation for gene function prediction in non-model *de novo* transcriptomes 106
4.1.1.1 Gene Ontology (GO) 107
4.1.1.2 Kyoto Encyclopedia of Genes and Genomes (KEGG) 108
4.1.1.3 The Enzyme Commission (EC number) 109
4.1.1.4 Enrichment analysis 109
4.1.2 DAVID enrichment analysis 110
4.1.3 Alkaloid biosynthesis studies 112
4.1.3.1 P450 subfamilies linked to secondary metabolism 113
4.1.3.2 Methylenedeoxy-bridge formation 114
4.1.3.3 Phenol coupling 114
4.1.3.4 PSRs role in secondary metabolism 115
4.1.4 Well-studied alkaloid biosynthetic pathways in plants 117
4.1.4.1 Benzylisoquinoline alkaloids in Papaver somniferum 117
4.1.4.2 Capsaicinoids in Capsicum annuum 119
4.1.4.3 Rosmarinic acid (RA) in Lamiaceae herbs 121
4.1.5 Investigating the galanthamine pathway 123

4.2 Methodology 125
4.2.1 GO annotation 125
4.2.2 Creation of the GOI (gene of interest) database from genes involved in secondary metabolite production in selected medicinally important plants to search for orthologs in daffodils 127
4.2.3 BLASTx search against GOI database for the 454 and post TE removal Trinity assemblies for putative transcripts involved in alkaloid production 131
4.2.4 GO enrichment studies for the GOI predicted transcripts 131
4.2.4.1 DAVID prediction of enrichment, EC and KEGG annotation 131
4.2.5 The determination of non-synonymous SNPS within the putative transcripts- Prediction of ORF and non-synonymous SNPs 132
4.2.6 The determination of putative transcripts with significant differences in transcript levels between daffodil varieties 133
4.2.6.1 BitSeq 133
4.2.6.2 Confirmation of transcripts via RT-PCR and transcript level differences via RT-qPCR 133
4.2.6.3 Confirmation of predicted transcripts 133
4.2.6.4 Sanger sequencing to confirm SNPs 134
4.2.6.5 Real time qPCR of those transcripts predicted to have significant transcript level differences 134

4.3 Results 136
4.3.1 Whole transcriptome functional annotation 136
4.3.1.1 Functional categorization of the 454 assembly 136

4.3.1.2 Functional categorization of the Original Trinity assembly 136
4.3.1.3 Functional categorization of the Post-TE removal Trinity assembly 136
4.3.1.4 The assignment of GO terms to transcripts 141
4.3.2 BLASTx search against the GOI database for the 454 and pre/Post-TE removal Trinity assembled transcripts for putative transcripts involved in alkaloid production 143
4.3.3 Pathway analysis: the use of DAVID to predict enrichment from EC and KEGG annotation 143
4.3.3.1 DAVID analysis for gene enrichment 143
4.3.3.2 KEGG pathway annotation of GOI daffodil transcripts 147
4.3.3.3 EC annotation of GOI daffodil transcripts 152
4.3.4 The determination of non-synonymous SNPs within the putative GOI transcripts - Prediction of ORFs and non-synonymous SNPs, confirmation of transcripts via PCR and Sanger sequencing confirmation of SNPs. 153
4.3.4.1 4.4.5.1 HDA57HA01AOVJD 154
4.3.4.2 4.4.5.2 Daff88927 156
4.3.4.3 Non-synonymous SNP verification 159
4.3.5 The determination of significant differences in transcript levels for transcripts selected from the GOI database 159
4.3.5.1 BitSeq 159
4.3.5.2 Confirmation of levels via comparative qPCR 161

4.4 Discussion 165
4.4.1 GO annotation and categorization of the whole transcriptome 165
4.4.2 BLASTx results 166
4.4.3 DAVID analysis, EC and KEGG annotation 166
4.4.4 Transcript level and SNP analysis 169
4.4.5 Transcripts linked to alkaloid biosynthesis 169
4.4.5.1 PAL and TYDC 169
4.4.5.2 CYP450s 171
4.4.5.3 HDA57HA0AK3FX 171
4.4.5.4 Daff88927 171
4.4.5.5 Comp75950_c0_s1 172
4.4.5.6 Comp97312_c0_s1 and Comp99544_c0_s1 172
4.4.5.7 Other transcripts of interest 172

5 CHAPTER 5 – OVERALL CONCLUSIONS AND FUTURE WORK 173

5.1 Transcriptomic reference for the daffodil 173

5.2 The uses of currently available Ontologies and Functional Annotations –bias towards model organisms 173

5.3 Determination of ploidy level of Carlton and Andrew's choice 174

5.4 The effect of polyploidism on gene expression studies 175

5.5 Prediction of genes involved in galanthamine production 176
5.5.1 PAL and TYDC 176
5.5.2 Comp75950_c0_s1 a homolog of C4H in vanillin biosynthesis as a putative transcript involved in the biosynthesis of protocatechuic acid in galanthamine biosynthesis 176
5.5.3 Contig01404 as a homolog for NCS in the biosynthesis of norbelladine 176
5.5.4 Daff106212 as a homolog for OMTs could be involved in the *O*-methyltransferase catalysed step of norbelladine to 4-*O*-methylnorbelladine 177
5.5.5 HDA57HA0AK3FX and Daff88927 as possible transcripts involved in C-C intermolecular phenol coupling in the final step of galanthamine biosynthesis 177

5.6 Predicted pathway for galanthamine biosynthesis 178

5.7 Future Work 180

5.7.1 Expression of candidate genes and subsequent functional analysis on the protein ... 180
5.7.2 Microbial engineering to create a galanthamine cell factory ... 181
5.7.3 Sequencing of genomic DNA to identify regulatory DNA motifs ... 181
5.7.4 Further investigation into the biological system in daffodil ... 182
5.7.5 The development of bioinformatics tools to further evaluate ploidy level in plants ... 182
5.7.6 The use of SNPs as biomarkers for daffodil breeding linked to alkaloid production ... 182

5.8 Conclusion ... 183

6 APPENDIX ... 185

6.1 **List of perl scripts on appendix disc ... 185**

6.2 List of data files available on appendix disc ... 186

6.3 Galanthamine level data ... 187
6.3.1 Introduction ... 187
6.3.1.1 Experiment 1: Comparison of Andrew's choice and Carlton from three different sites ... 187
6.3.1.2 Experiment 2: Comparison of the two varieties in the flowering stage of development ... 187
6.3.2 Methodology ... 188
6.3.2.1 Galanthamine extraction for Experiment 1 ... 188
6.3.2.2 Galanthamine extraction for Experiment 2 ... 188
6.3.2.3 Chromatographic techniques ... 188
6.3.3 Results ... 189
6.3.3.1 Experiment 1 ... 189
6.3.3.2 Experiment 2 ... 189
6.3.4 Discussion ... 189

6.4 Chromosome counting for Carlton and Andrew's choice varieties of daffodil ... 190
6.4.1 Introduction ... 190
6.4.2 Method ... 190
6.4.2.1 Reagents ... 190
6.4.2.2 Pretreatment ... 190
6.4.2.3 Fixation ... 191
6.4.2.4 Hydrolysis ... 191
6.4.2.5 Slide preparation ... 191
6.4.3 Results and discussion ... 191
6.4.3.1 Carlton: ... 191
6.4.3.2 Andrew's choice: ... 191

6.5 Andrew's Choice RNA extraction trial results and discussion ... 193

6.6 List of Plants and Genes in GOI database ... 196

6.7 Primers for RT-PCR, Sanger sequencing and qPCR ... 199

6.8 qPCR Plate set up ... 200

6.9 ClustalO Alignment from UniProt Alignment program ... 201

6.10 Command line prompts ... 203
6.10.1 Trinity assembly and Soapdenovo-Trans assembly –(Section 3.3.1.6) ... 203
6.10.2 Mapping Analysis –(Section 3.3.1.7) ... 204
6.10.3 VarScan and pileup_parser.pl –(Section 3.3.2) ... 205
6.10.4 BitSeq Analysis – (section 3.3.2) ... 206
6.10.5 BMtagger –(Section 3.3.2) ... 207
6.10.6 TransposonPSI.pl –(Section 3.3.5.2) ... 207

6.10.7 BLAST against TREP- (Section 3.3.5.3) 207
6.10.8 Trinity re-assembly- (Section 3.3.6) 208
6.10.9 BLAST against genes of interest database- (Section 4.2.3) 208

6.11 References 209

List of Figures

Figure 1-1 Daffodil variety Carlton. 4
Figure 1-2 The nine distinct types of amaryllidaceae alkaloids 6
Figure 1-3 Guillou *et al* synthetic pathway of galanthamine (adapted from Guillou *et al.,* 2001). 10
Figure 1-4 Enzymatic synthesis of norbelladine derivatives as shown by Eichhorn in *Leucojum vernum* (adapted from Eichhorn., 1998*)*. 12
Figure 1-5 Intramolecular para-ortho coupling reaction catalysed by a microsomal cytochrome P450 enzyme.[66]. 12
Figure 1-6 Chemical structure of widely used alkaloids (adapted from Takos *et al.,* 2013). 21
Figure 2-1 Annotation pipeline using "full_annotation.pl" 36
Figure 2-2 Total RNA electropherogram for sample 1 37
Figure 2-3 Total RNA electropherogram for sample 2 37
Figure 2-4 Depleted sample 1 electropherogram 38
Figure 2-5 Fragmented sample 1 electropherogram 39
Figure 2-6 Fragmented sample 2 electropherogram 39
Figure 2-7 High sensitivity DNA assay electropherogram for cDNA library from sample 1. 40
Figure 2-8 High sensitivity DNA assay electropherogram for cDNA library from sample 2. 41
Figure 2-9 Pipeline to show the steps of the annotation of the joint assembly. 44
Figure 2-10 Pipeline to show the steps involved in the annotation of the full assembly. 45
Figure 3-1 Schematic of a BWT. 57
Figure 3-2 differences between polyploids and diploids that alters copy number prediction. 60
Figure 3-3 Example of SAMtools pileup output format. 63
Figure 3-4 Example of output from pileup_parser.pl script 64
Figure 3-5 Schematic of Illumina data analysis. 73
Figure 3-6 Total RNA electropherogram from Andrew's Choice sample 1 81
Figure 3-7 Total RNA electropherogram from Andrew's Choice sample 2 82
Figure 3-8 Total RNA electropherogram from Carlton sample 1. 82
Figure 3-9 Total RNA electropherogram from Carlton sample 2 83
Figure 3-10 cDNA library electropherogram for Andrew's Choice sample 1. 84
Figure 3-11 cDNA library electropherogram for Andrew's Choice sample 2. 84
Figure 3-12 cDNA library electropherogram for Carlton sample 1. 85
Figure 3-13 cDNA library electropherogram for Carlton sample 2. 85
Figure 3-14 Venn Diagram showing SNP cross over between varieties 96
Figure 4-1 A simplified methylenedeoxy-bridge formation. 114
Figure 4-2 STR catalyzing a PSR reaction.(adapted from Stockigt *et al.*, 2011) [201] 116
Figure 4-3 Biosynthesis of several BIA subtypes (Modified from Desgagne 2012). 117
Figure 4-4 Biosynthesis of capsaicin (Modified from Aza Gonzalez *et al.*, 2010) 119
Figure 4-5 Chemical structure of protocatechuic acid and vanillin intermediates in the production of galanthamine and capsaicin. 120
Figure 4-6 Possible pathway for the production of protocatechuic acid in galanthamine biosynthesis. 121
Figure 4-7 Biosynthesis of rosmarinic acid in Lamiaceae species (Modified from Xiao 2011)[20]. 122
Figure 4-8 Proposed biosynthetic pathway of galanthamine. 123
Figure 4-9 Workflow of data analysis. 125
Figure 4-10 Screen shot of UniProt web interface for the retrieval program. 126
Figure 4-11 The initial results page from the UniProt retrieval program 126
Figure 4-12 The final output needed to assign GO terms to the transcripts. 127
Figure 4-13 A screen shot of the DAVID web interface for the functional annotation tool. 132
Figure 4-14 Molecular function assignment for the UniProt GO analysis of the 454 assembly. 138
Figure 4-15 Molecular function assignment for the UniProt GO analysis for the Original Trinity assembly. 138
Figure 4-16 Molecular function assignment for the UniProt GO analysis of the Post-TE removal Trinity assembly... 138
Figure 4-17 Cellular Component assignment from the UniProt GO categorization for the 454 assembly. 139
Figure 4-18 Cellular component assignment from the UniProt GO analysis for the Original Trinity assembly. 139
Figure 4-19 Cellular Component assignment for the UniProt GO analysis of the Post-TE removal Trinity Assembly. 139
Figure 4-20 Biological Process assignment from the UniProt GO analysis for the 454 assembly. 140
Figure 4-21 Biological Process assignment for the UniProt GO analysis for the Original Trinity assembly 140
Figure 4-22 Biological Process assignment for the UniProt GO analysis of the Post-TE removal assembly 140
Figure 4-23 Graph to show fold change in transcript level seen in Andrew's Choice compared to Carlton 162
Figure 4-24 Amplification trace for Daff88927 162

Figure 4-25 Simplified Trace for Daff88927 showing one biological replication.. 163
Figure 5-1 Putative pathway for the biosynthesis of galanthamine in daffodils. 179
Figure 6-1 Phase contract microscope image of Andrew's Choice cells. 192

List of Tables

TABLE 1-1 THE STATE OF SECOND-GENERATION SEQUENCING IN 2009. 17
TABLE 2-1 ASSEMBLY COMPARISONS OF THE TWO SAMPLES........ 42
TABLE 2-2 COMPARISON OF THE TWO SAMPLES MAPPING TO JOINT ASSEMBLY........ 42
TABLE 2-3 MAPPING OF THE THREE SEPARATE SEQUENCING SAMPLES TO THE FULL ASSEMBLY........ 43
TABLE 3-1 METHODS USED IN RNA EXTRACTION TRIAL FOR ANDREW'S CHOICE........ 71
TABLE 3-2 SCRIPTSEQ INDEX PCR PRIMERS USED IN LIBRARY PREPARATION. 72
TABLE 3-3 TOTAL RNA NANO CHIP RESULTS FOR THE FOUR TOTAL RNA SAMPLES FOR CONSIDERATION FOR LIBRARY PREPARATION. ... 80
TABLE 3-4 CDNA LIBRARY FINAL ANALYSIS........ 83
TABLE 3-5 COMPARISON OF CARLTON ILLUMINA READS AFTER ADAPTOR TRIMMING, QUALITY FILTERING AND QUALITY TRIMMING. 87
TABLE 3-6 COMPARISON OF ANDREW'S CHOICE ILLUMINA READS AFTER ADAPTOR TRIMMING, QUALITY FILTERING AND QUALITY TRIMMING. 88
TABLE 3-7 COMPARISON OF *DE NOVO* ASSEMBLERS. 89
TABLE 3-8 COMPARISON OF MAPPING TO THE 454, TRINITY AND SOAPDENOVO-TRANS ASSEMBLIES........ 90
TABLE 3-13 COMPARISON OF TRINITY ASSEMBLIES OF THE THREE TE REMOVAL METHODS. 94
TABLE 3-14 COMPARISON OF RAW READS MAPPED BACK TO THE 454 REFERENCE AND THE TRINITY ILLUMINA *DE NOVO* ASSEMBLY....... 95
TABLE 3-16 COMPARISON OF ASSEMBLY ANNOTATION OF ORIGINAL AND POST-TE REMOVAL ASSEMBLIES. 97
TABLE 4-1 THE ENZYMES INVOLVED IN THE BIOSYNTHESIS OF CODEINE, MORPHINE AND SANGUINARINE THAT HAVE BEEN ISOLATED OR CHARACTERIZED. 118
TABLE 4-2 THE ENZYMES INVOLVED IN THE BIOSYNTHESIS OF CAPSAICIN. 120
TABLE 4-3 THE ENZYMES INVOLVED IN ROSMARINIC ACID BIOSYNTHESIS........ 122
TABLE 4-4 THE FOURTEEN ENZYMES OF INTEREST USED IN THE GOI DATABASE. 129
TABLE 4-5 GO ANNOTATION RESULTS FOR THE 454 ASSEMBLY. 142
TABLE 4-6 THE GO ANNOTATION RESULTS FOR THE ORIGINAL TRINITY ASSEMBLY........ 142
TABLE 4-7 THE GO ANNOTATION FOR THE POST-TE REMOVAL TRINITY ASSEMBLY. 142
TABLE 4-14 SUMMARY OF KEGG RESULTS FOR THE 454 DATA. 148
TABLE 4-15 KEGG SUMMARY FOR THE ORIGINAL TRINITY ASSEMBLY. 149
TABLE 4-16 KEGG SUMMARY FOR THE POST-TE REMOVAL ASSEMBLY. 151
TABLE 4-17 SUMMARY OF SNP MINING OF GOI TRANSCRIPTS. 159
TABLE 4-18 TRANSCRIPT LEVEL RESULT SUMMARY. 164
TABLE 4-19 COMBINED RESULTS FROM THE CLUSTALO ALIGNMENT IN UNIPROT WITH THE BLASTX RESULTS AGAINST THE GOI FOR DAFFODIL TRANSCRIPTS. 170
TABLE 6-1 PERL SCRIPTS........ 185
TABLE 6-2 DATA FILES ON APPENDIX DISC. 186
TABLE 6-3. COMPARISON OF LEVELS OF GALANTHAMINE FOR EXPERIMENT 6.3.2.1 APRIL SAMPLES........ 189
TABLE 6-4 COMPARISON OF LEVELS OF GALANTHAMINE FOR EXPERIMENT 6.3.2.1 MAY SAMPLES. 189
TABLE 6-5 COMPARISON OF LEVELS OF GALANTHAMINE FROM EXPERIMENT 6.3.2.2. 189
TABLE 6-6 RNA EXTRACTION TRIAL RESULTS. 194
TABLE 6-7 PLANTS AND GENES INCLUDED IN GOI DATABASE........ 196
TABLE 6-8 PRIMERS FOR RT-PCR, SANGER SEQUENCING AND QPCR. 199
TABLE 6-9 QPCR PLATE LAYOUT........ 200

Abstract

The Amaryllidaceae have characteristic isoquinoline alkaloids including galanthamine that is approved for treatment of Alzheimer's disease. The daffodil (*Narcissus pseudonarcissus*) is an industrial source of this alkaloid. This project undertook analysis of the daffodil transcriptome as an approach to understanding this alkaloid biosynthetic pathway.

Material from the basal plate of var. Carlton was analysed using the Roche 454 GS FLX Titanium and Illumina HiSeq platforms to assemble reference transcripts (45324 transcripts from 454, 165065 from Illumina). Annotation was via a bespoke BLAST pipeline utilizing UniProt, TAIR, Rfam and RefSeq. Further functional annotation and enrichment studies were carried out using the DAVID platform encompassing KEGG, GO and EC annotations. Illumina HiSeq sequencing of a second variety, Andrew's Choice, was used alongside the reference transcripts to identify SNPs and transcript level differences. A bioinformatics method to determine ploidy indicated both varieties were triploid, in agreement with microscopy results. The level of selected transcripts was also assessed using qPCR.

Several transcripts putatively involved in alkaloid biosynthesis were identified. Comp75950_c0_s1 showed homology to a C4H gene from peppers and could be involve in protocatechuic acid biosynthesis in daffodils. Two transcripts, Daff106212 and Contig1404, were predicted to catalyse the synthesis of norbelladine from protocatechuic acid and tyramine, and its subsequence conversion to 4'-*O*-methylnorbelladine. Finally, transcripts HDA57HA0AK3FX and Daff88927 were suggested for the final step in galanthamine biosynthesis, an intermolecular phenol coupling.

This is the first transcriptomic comparison of two daffodil varieties and is an important resource for further investigation into genes involved in Amaryllidaceae alkaloid biosynthesis.

Abbreviations

%	Percentage
°C	Degree Centigrade
4CL	4-coumaroyl-CoA ligase
AChe	Acetyl-choline esterase
ACL	Acyl carrier protein
ACS	Acyl-CoA synthetase
AFLPs	Amplified Fragment Length Polymorphisms
AGI	Arabidopsis Genome Initiative
ATC	Chloroplast material
ATM	Mitochondrial material
ATP	Adenosine triphosphate
BCAT	Branched-chain amino acid transferase
BIA	Benzylisoquinoline alkaloids
BitSeq	Bayesian Inference of Transcripts from Sequencing Data
BLAST	Basic Local Alignment Search Tool
BMTagger	Best Match Tagger
bp	Base pair
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transformation
C3H	Coumaroyl shikimate/quinate 3-hydroxylase
C4H	Cinnamate 4-hydroxylase
CCoAOMT	Caffeoyl-CoA 3-*O*-methyltransferase
cDNA	Complementary DNA
CGR	Centre for Genomic Research, University of Liverpool
CheSyn	Cheilanthifoline synthase
ChIP-Seq	Chromatin immunoprecipitation sequencing
cm	Centimetre
CMBR	Contigs mapped back to reference
CODM	Codeine O-demethylase
COR	Codeinone reductase
CPU	Central Processing Unit
CS	Capsaicin or capsaicinoid synthase
CT	Threshold cycle
CTAB	Cetyltrimethylammonium bromide
DAVID	Database for Annotation, Visualization and Integrated Discovery
DE	Differential expression
DNA	Deoxyribonucleic acid
dNTPS	Deoxynucleotide triphosphate
EASE	Expression Analysis Systematic Explorer
EC	Enzyme Commission
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed Sequence Tag

FAT	Acyl-ACP thioesterase
Fe	Iron
FM-index	Ferragina Manzini index
FU	Fluorescence Unit
g	Gram
g	Gravitational force
$g\ l^{-1}$	Gram per litre
GB	Giga bytes
Gbp	Giga base pairs
GO	Gene Ontology
GOI	Genes of Interest
GSEA	Gene set enrichment analysis
HCHL	Hydroxycinnamoyl-CoA hydratase/lyase
HCT	Hydroxycinnamoyl transferase
HPPR	Hydroxyphenylpyruvate reductase
ID	Identity
Indels	Insertions and Deletions
ITMI	International Triticeae Mapping Initiative
KAS	Ketoacyl-ACP synthase
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LiCl	Lithium chloride
LTR	Long Terminal Repeat
M	Molar
Mbp	Mega base pairs
MCMC	Markov chain Monte Carlo
MEA	Modular Enrichment Analysis
MIA	Monoterpene indole alkaloids
MID	Multiplex Identifier
min	Minute
MGI	Mouse Genome Informatics
ml	Millilitre
mm	Millimetre
mM	Millimolar
mRNA	Messenger RNA
NaCl	Sodium chloride
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Centre for Biotechnology Information
NCS	Norcoclaurine synthase
NICE	The National Institute for Health and Care Excellence
ng	Nanograms
NMT	*N*-methyltransferase
NMCH	(S)-N-methylcoclaurine 3'-hydroxylase
nt	Nucleotide

OD	Optical Density
OMT	*O*-methyltransferase
ORF	Open Reading Frame
P6H	Protopine 6-hydroxylase
PAL	Phenylalanine ammonia-lyase
PCR	Polymerase Chain Reaction
pg	Pico grams
PPLR	Probability of Positive Log Ratio
PSR	Pictet Spenglerase
PVP40	Polyvinylpyrrolidone
qPCR	Quantitative PCR
RAPDs	Random Amplified Polymorphic DNA
RAS	Rosmarinic acid synthase
RDMs	Random markers
RefSeq	NCBI Reference Sequence database
Rfam	RNA family database
RFLPs	Restriction fragment length polymorphism
RIN	RNA Integrity Number
RL	Rapid Library
RMBT	Reads mapped back to transcripts
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RT-PCR	Real Time PCR
SalAT	Salutaridinol 7-O-acetyltransferase
SalR	Salutaridine reductase
SalSyn	Salutaridine synthase
SEA	Singular Enrichment Analysis
SGD	*Saccharomyces* Genome Database
SMRT	Single molecule real time
SNP	Single Nucleotide Polymorphism
SSRs	**Simple sequence repeats**
STR	Strictosidine synthase
T6ODM	Thebaine 6-O-demethylase
TAIR	The Arabidopsis Information Resource
TB	Terabyte
TE	Transposable element
TNMT	Tetrahydroprotoberberine N-methyltransferase
TREP	Triticeae Repeat Sequence Database
Tris-HCl	Tris-hydrochloride
TYDC	Tyrosine decarboxylase
TyrAT	Tyramine aminotransferase
µg	Microgram
µl	Microlitre

1 Chapter one: The use of Second-generation sequencing technologies for the study of medicinally important alkaloids - a case study in galanthamine production in daffodils

1.1 Introduction

Galanthamine is a specialized plant secondary metabolite currently extracted from snowdrops and daffodils for use in the pharmaceutical industry as the active ingredient in a drug that slows the progression of Alzheimer's disease [1]. Plant secondary metabolites such as this have been exploited throughout history as flavours, pigments, medicines and as industrial raw materials [2]. The compounds themselves are often species-specific and are thought to be involved in plant defense and in the attraction of pollinators [2]. Galanthamine is a member of a group of metabolites known as alkaloids that also includes other compounds with potent biological activity such as codeine and morphine [3]. The plants often remain the main source of these compounds as the chemical synthesis is difficult due to the chiral nature of the compounds [4]. The direct extraction from plants, however still produces low yields as the levels of alkaloids in plants vary greatly between species and individuals as well as in response to environmental conditions and therefore they are considered trace compounds [5]. Although methods of metabolic engineering and biosynthesis via microbial systems have been successful for compounds of this type, these methods rely heavily on knowledge on the biosynthesis of the compounds within the source species [6].

As secondary metabolite biosynthesis often involves diversions from primary metabolism and complex non-linear pathways it is difficult to predict the enzymes involved [7]. Although studies have shown that the enzymes involved come from a small number of gene families, these families have large variation of function within them and it is often not possible to predict function using sequence analysis alone [8].

One method of investigating the biosynthesis of alkaloids is functional genomics. Major advances in DNA sequencing technology have made genomic sequencing a relatively cheap and quick way of obtaining vast quantities of data, even from non-model plants [9-11]. With the rapid development and corresponding reduction in cost of second-generation sequencing technologies, transcriptomics has become the go-to method of functional genomic studies in previously unexploited non-model plants [12,13]. In particular the massively parallel sequencing of RNA, "RNA-Seq", has revolutionized plant research, from the initial uses of transcriptomics for EST library production, mapping of short reads to reference genomes to the development of *de novo* assembly of the improved read lengths of second-generation techniques, transcriptomics has led the way to the discovery of genes involved in alkaloid biosynthesis in several plants [5,14,15].

Two key transcriptomic analyses in the discovery of putative alkaloid biosynthetic genes are similarity searches and functional annotation of the transcriptome. Functional annotation via BLAST searches against public databases of Gene Ontology, KEGG and EC classification have been used to predict genes involved in other alkaloid producing pathways [2,5,16]. Also, with the enzymes involved in alkaloid production predicted to be from a small number of large gene families, homology studies into specific gene types has been successful in predicting alkaloid biosynthetic genes in some plants [17,18]. These predicted genes can then be used in further downstream analysis to aid in the discovery of plants and systems with higher alkaloid producing properties.

One valuable tool in the analysis of these predicted genes is transcript level differences or differential expression (DE) [13,16,19,20]. By combining DE studies with transcriptome sequencing it is possible to assess key steps in alkaloid production [21]. As a transcriptome shows all the transcribed elements (exonic regions) at a given time point, comparisons of daffodil individuals with differing levels of galanthamine from the same time point in growth may predict genes that are up or down regulated within the galanthamine pathway.

The rapid development of functional genomics has also lead to the revolution of DNA markers used in plant research and breeding programs [22]. The first and second generation of markers such as RFLPs, RAPDs, SSRs and AFLPS rely on costly gel based assays and are time consuming [22]. The development of SNPs, a third generation marker system, not only avoids the gel-based methods but SNPs are also considered functional markers, linking traits with alleles[23]. They are the most abundant marker system and have the potential to lead to agronomically important alleles [24]. SNPs have been used in agriculturally and medicinally important plants as cost effective marker-associated selection in fingerprinting, association studies and population analysis [25]. In this project it is hoped that by discovering SNPs between two varieties of daffodil with known differences in galanthamine levels (see appendix section 6.3) it will be possible to predict SNP markers linked to galanthamine production.

This project aims to produce a reference transcriptome of a high galanthamine producing variety of daffodil, Carlton, at a time point known to correlate with high galanthamine levels. By comparing this reference to short reads from both Carlton and a low producing variety, Andrew's Choice, it is hoped that both non-synonymous SNPs and differences in transcript levels can be determined in putative genes involved in the production of galanthamine.

1.2 Daffodil- Narcissus pseudonarcissus

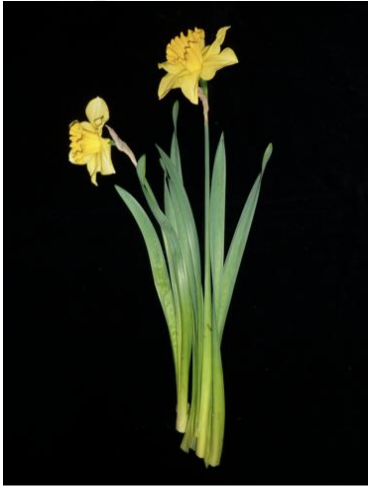


Figure 1-1 Daffodil variety Carlton.

The *Narcissus* genus is among the 80 genera of the monocotyledon Amaryllidaceae family, and includes all daffodils [26]. They originate from the Mediterranean area, from lowland pastures to rocky hillsides [27,28]. The plants undergo hybridization easily in both the wild and cultivation, resulting in over 25,000 distinct cultivars [26]. All species are summer-dormant bulbs that grow throughout the autumn/winter with most flowering in the spring, and are insect pollinated [28].

The taxonomy of the family is highly contested with *Narcissus pseudonarcissus* being one of the most controversial groups due to the ease of hybridization and varying DNA content and polyploidism [29]. The genus has been shown through several cytogenetic and flow cytometry studies to have diverse DNA content and ploidy levels, with genome size and ploidy differing within seven of the most studied species, (*N. asturiensis*, *N. bulbocodium*, *N. broussonetii*, *N. cantabricus*, *N. poeticus, N. pseudonarcissus* and *N. tazetta*) [30]. One of the most extensive cytometry studies is the 2008 study of 355 accessions by Zonneveld. He identified highly variable DNA content within varieties with the same ploidy level. For example, in diploids the somatic nuclear DNA content (2C) varied from 14 to 38 pg [30]. As for polyploids the largest polyploid reported was the nonoploid *N. dubius*, which had a DNA content of 96.3 pg.

The varieties studied in this project are *N. pseudonarcissus* cv. Carlton and *N.*

pseudonarcissus cv Andrew's Choice. In his study Zonneveld did not look at either of these varieties, although he did measure the DNA content of 46 *Narcissus pseudonarcissus* samples with DNA content ranging from a diploid (2n=14) ssp *pseudonarcissus* L with 22.7pg to a tetraploid (2n=28) ssp *nobilis* with a DNA content of 45.9pg [30]. The majority of the samples were diploid apart from *N nobilis* and two triploid samples of *ssp major* (Curtis) Baker also known as *N. hispanicus* that had DNA contents of 36.5 and 36.1pg respectively [30]. Carlton was investigated in 1993 in a study looking at only 16 varieties of *Narcissus* and was found to be a tetraploid; Andrews Choice has not been investigated for DNA content or ploidy level [31]. This diversity within the genus has been exploited in the ornamental industry to produce a wide variety of interesting phenotypes [32]. As well as interest from the ornamental breeders daffodils have become an important medicinal plant due to their production of biologically active alkaloids such a galanthamine [33].

1.3 Amaryllidaceae alkaloids – History of their medicinal properties

There are 500 structurally diverse Amaryllidaceae alkaloids which have been identified by progressive chemical analysis since the 1950s [34]. Within the *Narcissus* genus alone there are over 100 known alkaloids with varying bioactivity [35]. The earliest record of the use of galanthamine, the most widely studied Amaryllidaceae alkaloid, is possibly from ancient Greece. It has been suggested that symptoms matching that of acetylcholine syndrome are described by Homer in the Odyssey and were alleviated by a plant extract known as "moly" thought to be from the snowdrop (*Galanthus spp.*) which was known to grow in Greece [36]. Snowdrops have been used in Russia and Eastern Europe for centuries in traditional medicine. Although their endogenous role is relatively unknown, extracts of Narcissus have been used for centuries to treat a wide variety of ailments [36]. Through second hand accounts and unconfirmed reports extracts have been linked to the treatment of post-polio paralysis and myasthenia gravis via the reversal of neuromuscular blockade, as well the ease of nerve pain and other neuromuscular or central nervous system disturbances [1].

Galanthamine is a member of one of the nine distinct groups of Amaryllidaceae alkaloids [37]. These are grouped based on their skeletal characterization, as shown in Figure 1.2. The bioactivities of these alkaloids are varied but include analgesic properties, acetylcholine esterase inhibition, hypotensive properties, anticonvulsive properties, anti-inflammatory properties, cytotoxic properties and antimalarial properties [37].

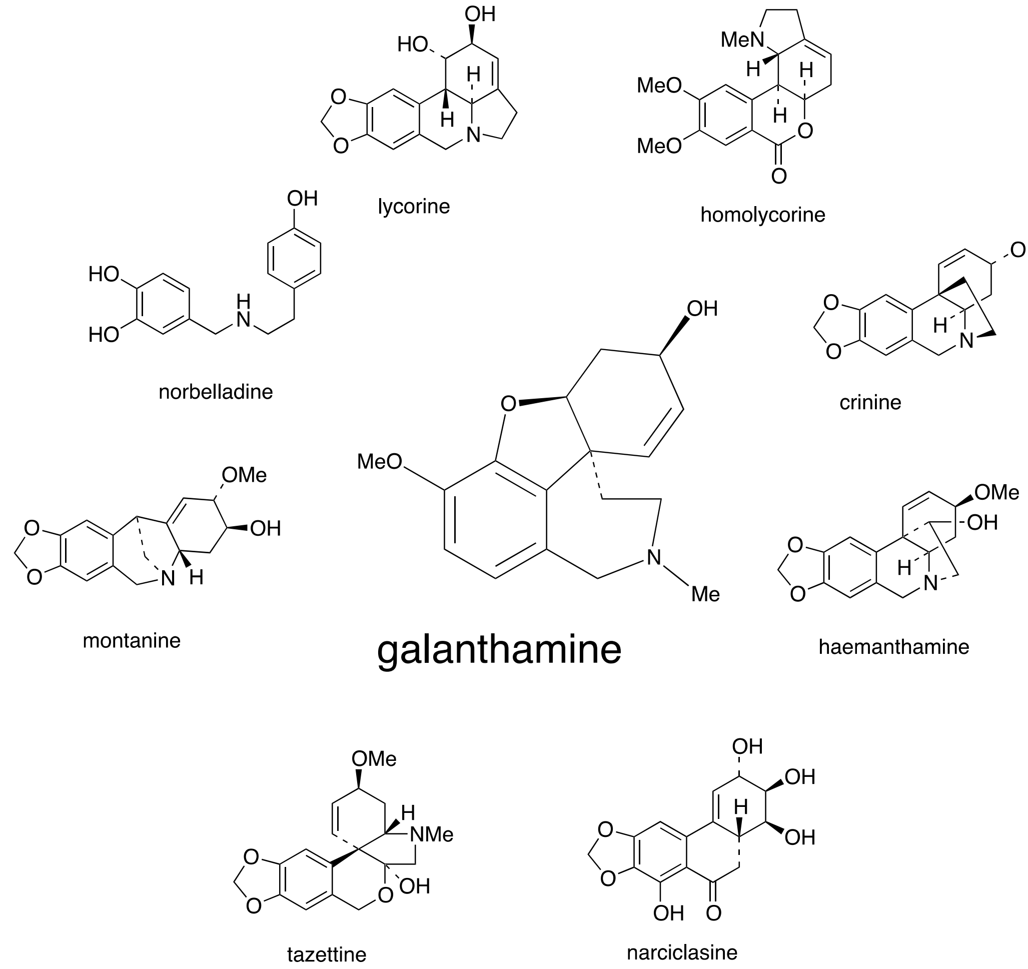


Figure 1-2 The nine distinct types of amaryllidaceae alkaloids

The alkaloids shown are the representative alkaloid for that type [37].

1.3.1 Galanthamine – as a folk medicine

Galanthamine is the most widely studied Amaryllidaceae alkaloid and is currently approved for use in the treatment of Alzheimer's disease and is the main focus of this thesis [38]. In the 1950s the active ingredient galanthamine was first characterized after a Bulgarian pharmacologist witnessed snowdrop extracts being used to alleviate headaches [1]. The structure was determined in 1953 and its acetylcholine esterase inhibiting properties discovered shortly thereafter [39,40]. It was also implicated as a treatment for poliomyelitis in Eastern Europe in the 1960s when extracts from Caucasian snowdrop bulbs were effective on two children showing early symptoms [1]. It is however because of its acetylcholine esterase (AChE) inhibiting properties that galanthamine is of pharmacological interest.

1.3.2 Other Amaryllidaceae alkaloids with potential medical uses: Narciclasine and Lycorine

Most of the work on amaryllidaceae alkaloids has focused on the chemical structure of the pharmacophores and functional groups to look for possible new lead compounds for structural based drug design. Lycorine has shown activity against several cancer lines as it is shown to induce apoptosis via the mitochondria of cancerous cells [41,42]. This is of particular interest as lycorine shows cytostatic activity in those cells resistant to apoptosis, since 90% of cancer patients die from metastases, which are intrinsically resistant to apoptosis [43]. Derivatives of lycorine have shown anti-plasmodial action against sleeping sickness, malaria, and also against the viral diseases poliomyelitis and SARS [44-46]. Derivatives of this alkaloid are also currently being explored for anti-dengue virus activity and as a broad spectrum anti pathogenic fungi agent with a study looking at 24 crop pathogenic fungi [47,48].

Narciclasine has also shown anti-cancerous properties against those cells resistant to apoptosis stimuli. It is thought to target cEFIA elongation factor causing cyto-skeletal disorganization [49]. More recently it has been shown to impair actin cytoskeleton organization in experimental models of brain cancers [50].

1.4 Galanthamine

1.4.1 As a proven drug in Alzheimer's disease

Galanthamine is a centrally acting competitive and reversible ACHE inhibitor approved by the National Institute for Health and Clinical Excellence (NICE) for the treatment of mild to moderately severe cases of Alzheimer's disease [38,51]. It is currently extracted for this purpose from snowdrops (*Galanthus spp*), red spider lily (*Lycorus radiate*) and daffodils (*Narcissus spp*) with varying intra and inter-species yields [52].

Alzheimer's disease is the main cause of dementia in the elderly with over 20 million sufferers worldwide; in 2009 there were 500,000 sufferers in the UK alone, reaching almost 800,000 by 2012 [38,51,53]. There is no known cure for the disease and due to prolonged life expectancy the worldwide cost is predicted to double every five years with the worldwide cost in 2005 alone estimated at US$315 billion. Within the UK, in 2012 it was predicted to cost £23 billion a year with a predicted increase to £27 billion by 2018 [53,54]. Alzheimer's disease is a progressive neurodegenerative disease resulting in serious cognitive dysfunction with symptoms including memory loss, language deficits, depression, behavioral issues, psychosis and can lead to motor dysfunction and Parkinson-like symptoms [52,55]. The pathology of the disease is relatively unknown but the major causative factors are known to be deficiency of acetylcholine (Ach), a neurotransmitter linked to cognitive function, along with plaque build-up and inflammation in the brain [55]. Therefore it is very important that drugs used to treat the disease, like galanthamine, can pass through the blood brain barrier.

Galanthamine has a dual mechanism against Alzheimer's disease. Firstly it has been shown to inhibit the enzymatic activity of AChE in the brain resulting in increased levels of ACh and secondly it allosterically modulates nicotinic ACh receptors, increasing the stimulatory effect of ACh [1]. This second mode of action is attributed to galanthamine's ability to bind at both the pre and post-synaptic nicotinic receptors at a different site to ACh allowing the response to be increased as both bind simultaneously [56,57]. This dual action makes galanthamine a more attractive treatment than other currently available drugs such as donepezil and rivastigmine as they do not affect the nicotinic receptors [35]. Due to the complex structure of galanthamine the main source of the compound remains plants as chemical synthesis produces very low yields at high cost.

1.4.2 Chemical synthesis

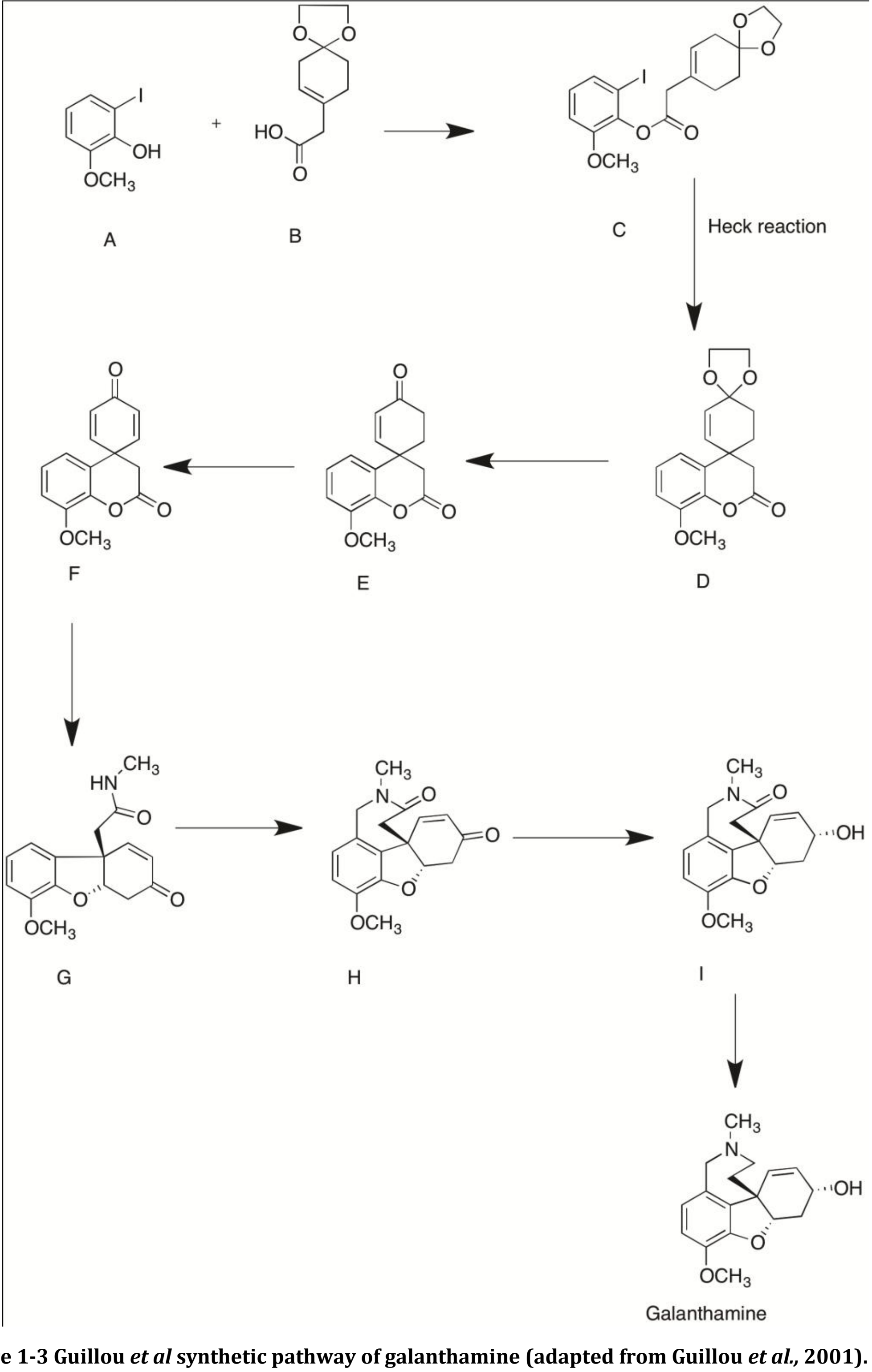


Figure 1-3 Guillou *et al* synthetic pathway of galanthamine (adapted from Guillou *et al.,* 2001).

Synthesis was achieved via oxonarwedine (H) in an eight-step Heck reaction leading to a 12% yield of galanthamine [58].

Total synthesis of galanthamine is possible but it is not commercially viable due to the intrinsic complexity of the chiral centres. The main limiting step is the unfavorable intramolecular oxidative para-ortho coupling of the phenolic ring [4,59]. Barton and Kirby carried out the initial chemical synthetic work on galanthamine and in 1962 they successfully synthesized galanthamine. This was achieved at a 1% yield via biomimetic and intramolecular phenol coupling to simultaneously give the quaternary carbon centre and tetracyclic framework [60]. In more recent years the yield has been improved to levels of about 12% by Guillou *et al* in 2001 and Tanimoto *et al* in 2007. Guillou based the synthesis on that of +- oxonarwedine using an eight-step process (see Fig 1.3) involving a Heck reaction resulting in a 12% yield [58]. Tanimoto *et al* utilized an 11-step process starting from D glucose to give a yield of 12.8%. The stereo specific quaternary carbon is created using a Claisen rearrangement on the chiral cyclohexanol derived from the D-glucose [61]. The low yields and complicated synthesis of galanthamine results in the continued use of the plant for commercial production.

1.4.3 Biosynthesis

As is proven with other alkaloids, studying the biosynthetic pathway can lead to valuable information that can be used to increase biotechnological production [62]. It could eventually lead to the cloning and expression of the rate limiting enzymes responsible for phenol oxidative coupling reaction [4]. Limited work has been carried out to try and elude the biosynthetic pathway of galanthamine. Barton and Cohen in 1957 suggested that all amaryllidaceae alkaloids are derivatives of norbelladine via intra-molecular oxidative phenol coupling after a radiolabelling experiment using α ^{14}C-labelled norbelladine derivatives as precursors in *N. pseudonarcissus cv.* King Alfred. [63]. It was predicted that a dienone was the first intermediate but the ether bridge formation mechanism was unknown. In 1969 work by Fuganti suggested the precursor was 4'-O-methylnorbelladine (R = CH_3 in figure 1.4) via the incorporation of this compound into galanthamine while studying the biosynthesis of haemanthamine [64]. In 1970 Bhandark and Kirby suggested it was narwedine (R=H in figure 1.4) following an experiment that incorporated ^{3}H narwedine into galanthamine [64,65]. The work carried out by Eichhorn set out to test these theories as the previous experiments had poor incorporation rates. He used carrier free radioactive labeled precursors to achieve up to 27% incorporation rate, proving 4’-O-methylnorbelladine to be the primary universal precursor [4].

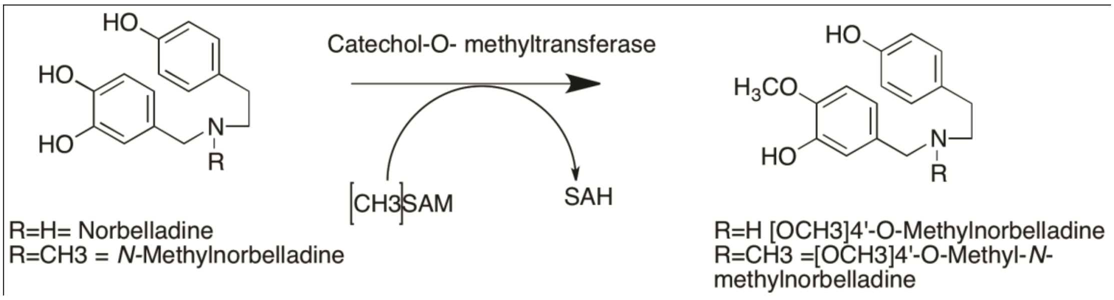

Catechol-O- methyltransferase
[CH3]SAM
SAH
R=H= Norbelladine
R=CH3 = *N*-Methylnorbelladine
R=H [OCH3]4'-O-Methylnorbelladine
R=CH3 =[OCH3]4'-O-Methyl-*N*-methylnorbelladine

Figure 1-4 Enzymatic synthesis of norbelladine derivatives as shown by Eichhorn in *Leucojum vernum* (adapted from Eichhorn., 1998*).*

The experiments showed that *N*-methylnorbelladine was the primary precursor [4].

The key step in galanthamine production is the phenol oxidative para-ortho’ coupling of 4’-O-methylnorbelladine to yield a hypothetical dienone. These reactions involve the oxidation of phenols by one-electron transfers, producing radicals that pair to form new C-C or C-O bonds [66]. From other studies on similar pathways it can be predicted that this step is carried out via a highly specific P450 dependent oxidase without introducing oxygen in the final product [4,67,68]. P450s are a class of haem protein-dependent mixed function oxidase that use NADPH or NADH to produce an organic compound and a molecule of water [69]. In BIA (Benzylisoquinoline alkaloids) production a very similar step involving intra-molecular para-ortho’ coupling of (R)-reticuline to the dienone salutaridine (Fig 1.5) is catalyzed via a P450 linked NADPH and O_2 dependent microsomal bound plant specific enzyme [70].

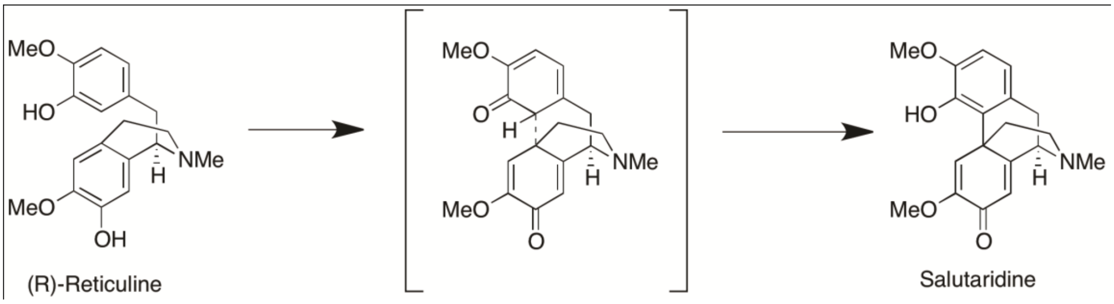


Figure 1-5 Intramolecular para-ortho coupling reaction catalysed by a microsomal cytochrome P450 enzyme.[66].

The role of P450 enzymes in phenol coupling reactions in alkaloid biosynthesis and other enzymes characterized in similar pathways is further discussed in chapter four.

1.5 The use of functional genomics in plant science

As discussed the rapid development of this area has lead to a cheap and efficient method of data collection on non-model plants such as daffodils. The main turning point came with the release of second-generation sequencing technologies.

1.5.1 The development of Second-generation sequencing technologies

Sanger sequencing was the most widely used DNA sequencing method from its widespread introduction in the 1980s until the early 21st century[71]. Within the plant community reference genomes were sequenced using this method, starting with the landmark Arabidopsis genome in 2000, providing the first plant genome sequence, followed by important crop plants such as rice, sorghum and soybean [71-75]. Sanger sequencing is time consuming, costly and requires detailed sample preparation. It often requires DNA cloning in bacteria which can result in host bias [72]. Due to the high cost and amount of time taken to produce the data, large and complex genomes of important plants such as wheat (~16GB genome size) could not be determined using this method. Furthermore, most plants with specialized metabolites of medicinal interest such as daffodils are of interest to small communities of researchers and therefore are likely to have few genomic resources [2]. New technologies that could produce more data at a lower cost were required for larger genomes and non-model organisms, resulting in development of a new generation of sequencing techniques known as “second-generation sequencing”.

Prior to the development of second-generation sequencing technologies the standard method for collecting genomic information on non-model plants was the generation of random expressed sequence tags (ESTs) via Sanger sequencing 73. Second-generation methods set out to confront the intrinsic problems associated with Sanger sequencing, by using a "sequencing by synthesis" approach as a quicker and cheaper process 74 that resulted in 3-4 times the magnitude of DNA sequence compared to Sanger, making it a viable method for understanding complex genomes and non-model organisms 71.

The key characteristics of second-generation technologies that made them fundamentally different were the introduction of *in vitro* cloning and *in situ* amplification as well as the use of chain extension chemistry as opposed to chain termination. The three main commercial products, 454-pyrosequencing, Illumina and Ion Torrent all share similar methodologies, with individual clonal DNA templates being sequenced in parallel via cycles of base additions and imaging [76]. They differ in their methods for detecting incorporation. Each platform is briefly described below.

1.5.1.1 454 pyrosequencing

Roche's 454 pyrosequencing is a "sequence by synthesis" method and is centred around pyrophosphate chemistry using beads containing sulfurylase and luciferase [77]. A single-stranded template DNA library is created via shearing the DNA into fragments that are ligated to two adaptors that allow the immobilization of the strands onto the beads [78]. The DNA fragments are amplified independently via emulsion-based amplification and sequencing occurs with one nucleotide being used per cycle [79]. The incorporation of this nucleotide results in the release of pyrophosphate that is converted via the sulfurylase to ATP. It is the hydrolysis of this ATP by luciferase, releasing oxyluciferin and light, that is measured followed by a wash and the next cycle [77]. The 454 chemistry results in indel sequencing errors, in part due to the lack of terminating moieties resulting in multiple incorporations in any one cycle [77]. This is particularly evident in sequences with regions of homopolymers as it is difficult to distinguish between high numbers of the same nucleotide such as 4

As instead of 5 As.

1.5.1.2 Illumina sequencing

Illumina sequencing technology (Illumina Inc.) is also a “sequencing by synthesis” method. The DNA in this case is also amplified *in situ* via bridge PCR to result in clusters of around 1000 copies [77]. Unlike other methods Illumina uses nucleotides labeled both with a fluorescent dye and a terminating moiety allowing for the use of all four nucleotides at once [80]. After a nucleotide is incorporated it is detected and identified according to its dye, the dye and termination group is removed and the next cycle occurs [77]. Due to the use of all four nucleotides at once and the termination group preventing numerous incorporations per cycle Illumina does not suffer the same indel sequencing errors as seen in 454 and Ion torrent. The main error in Illumina sequencing is substitution, that is identification of the wrong nucleotide caused by incomplete or missing blocking groups or residual interference from incomplete cleavage of the fluorescent label in previous cycles [81]. Illumina sequencing and its comparison to 454 are further explained in chapters two and three.

1.5.1.3 Ion Torrent

Ion Torrent (Life Technologies Corporation) was first implemented in 2010, and is a semiconductor sequencing technology. Unlike the 454 and Illumina methods it does not rely on enzymatic reactions, fluorescence or chemical luminescence [82]. It exploits the biochemical reactions that occur during nucleotide incorporation leading to the release of hydrogen ions that in turn cause a change in pH [82]. The technology involves the use of a micro-array chip which is flooded with one nucleotide at a time. If the nucleotide is incorporated a change in pH is recorded via an ion-sensitive layer under the wells [10]. This method was intended to be high speed and relative low cost with a run time of around two hours and a capacity for >1GB of data in 2012 [82]. However, as of 2011 its read length capabilities were only around 200bp and so it was considered more suitable for use in microbial studies, re-sequencing and was not widely used in plant transcriptome projects [82]. A further limitation to Ion

Torrent was its poor ability to handle homopolymers. The error associated with Ion Torrent was predominately indel based (error rate of 1% per base) [81]. However, unlike the other methods, its use of native nucleotides avoided the noise seen in both 454 and Illumina from the use of fluorescence or blocking groups.

The second-generation technologies of Roche 454 (e.g. GS FLX Titanium) and the Illumina platforms (such as HiSeq) were therefore considered high-throughput techniques [75,76] appropriate for non-model plant genomic studies and from 2010 to 2013 were the most widely adopted methods. The Roche 454 produced over a million reads at lengths of up to 700 bp in a 10 hour run with total amounts of 400-600 megabases of sequencing. Illumina HiSeq 2000 had a raw base accuracy of over 99.5%, producing on average over 1 billion high quality reads during an 11 day run, resulting in gigabites of data [71].

1.5.2 From genome to transcriptome

One of the most discussed and studied areas surrounding second-generation sequencing is its use in transcriptomics, otherwise known as "RNA-Seq" [11,83,84]. RNA-Seq involves the sequencing of the transcribed regions of DNA (mRNA that is converted to cDNA). Through focused studies on the coding regions the amount of repetitive regions sampled within the sequence data is reduced, increasing the informative content and easing assembly. It is possible using this method to look at the complete repertoire of transcribed events occurring in a specific tissue at any one time and is therefore of particular use in non-model large genome plants with limited genomic data. This method can be used to characterize genes, look for novel transcripts, compare mutations and gene expression between individuals and produce *de-novo* assemblies of reference transcriptomes [84].

1.5.3 The use of second-generation sequencing in plant science

Since the introduction of second-generation sequencing techniques in 2005, research into non-model plants with large genomes has developed rapidly. Transcriptomic projects are becoming the norm for research into non-model plant species. The first commercially available platform, Roche's 454 pyro-sequencing was in 2010/2011 the system of choice for *de novo* assembly of transcriptomes due to its longer read lengths as can be seen in Table 1.1 [11].

Table 1-1 The state of second-generation sequencing in 2009.

The three commercially available technologies vary greatly in run time, read length and number. (All data taken from Deschamps and Campbell, 2009) [25]

Sequencing platform	**Run time**	**Read length (bp)**	**Reads per run (million)**
Roche 454 FLX	10 hours	400-500	~1
Illumina GAIIx	5.5 days	100	160
ABI SOLiD	6-7 days	50	500

When these platforms first emerged *de novo* assembly was thought to be impossible due to the short read lengths first achieved by these technologies (in 2008 454 reached read lengths approaching 300bp whereas Illumina and SOLiD were close to 35bp) [85-88]. The data from the sequencers of this new generation were therefore originally used for sequence consensus applications, gene expression analysis, genome annotation, EST library production, discovery of small RNA molecules and SNP profiling [77,89]. As of the end of February 2010, four of the seven hundred and forty eukaryotic genome projects involved second-generation sequencing [77]. Other second-generation projects looking at plants included 454 EST library constructions for *Arabidopsis thaliana, Medicago truncatula* (a model legume) and *Zea mays* (maize). However, following the first fully second-generation *de novo* genome assembly of the giant panda in 2010, the use of second-generation sequencing for *de novo* projects has rapidly increased [90].

De novo transcript assembly has become a rapid and viable approach for plants with large genomes and 454 rapidly became the technique of choice for *de novo* assembly, with Illumina's short reads being used alongside these reference transcriptomes for SNP discovery and transcript differences to investigate numerous biological pathways [77]. This was in part due to the increased coverage seen with Illumina over 454, due to its capacity to produce 100 fold more reads with a 5 fold increased read depth on assembled contigs [2]. With the introduction of the Illumina HiSeq2000 in mid 2010 with a 2-5 fold rate increase in data acquisition over the GA series, Illumina sequencing became the choice for short read mapping back to references [91]. The increased coverage seen with Illumina compared to 454 allows for investigation into rare transcripts.

1.5.4 Transcriptomic studies on non-model plants

Between 2007, with the proof of concept transcriptome study of *Arabidopsis thaliana* seedlings, and 2012, there were over 50 plant transcriptomic studies involving RNA-Seq via either 454 or Illumina [13,83,92]. As read lengths improved (454 in 2005 100bp to 700bp in 2014, Illumina 25bp in 2006 to 300 in 2013) the second-generation technologies overtook traditional sequencing methods for a variety of studies including EST library creation, *de novo* transcriptome, genome annotation, discovery of markers (SNPs and SSRs) and a wide variety of studies looking as specific traits and relationships [13,83]. *De novo* transcriptome assembly has been carried out on non-model plants including certain species of fern and eucalyptus as well as garlic, pea, chestnut and chickpea [93-98], as well as being used in larger collaborative efforts such as the 1KP project, aiming to sequence over 1000 plants. To date this latter project has generated data on over 1300 samples of which 111 are monocots and 6 from the Amaryllidaceae family [99]. So far the project has not released its main publication but a list of companion papers can be found at:

https://pods.iplantcollaborative.org/wiki/display/iptol/OneKP+companion+papers [99].

As well as *de novo* assembly, second-generation sequencing has been used to analyse primary and secondary metabolism, traits such as C4 photosynthesis and response to biotic and abiotic stress [15,96,97,100-103]. Even with the rapidly growing number of transcriptomic projects on plant species only a fraction of the plant world has so far been explored.

1.5.5 The future of sequencing technologies

Study of transcriptomics in plants is continuing to grow with the ever-changing state of sequencing technologies. A new era of third-generation technologies has now been ushered in. Second-generation technologies such as 454 that were at the forefront of non-model transcriptome studies only a few years ago are now giving way to the newer technologies with the growing desire for longer reads, reduced computational requirements and cheaper sequencing driving the technology forward. In fact Roche has decided to no longer support the 454 platform and it is set for full decommission by mid 2016 [104]. The development of the new generation of sequencers is, as in the past with the second generation platforms, being pushed forward by human genetics with the aim of producing a whole genome sequence for less than $1000 [9].

The third-generation sequencers include the commercially available SMRT (single molecule real time) sequencer by Pacific Biosciences and Oxford's nanopore sequencers [105,106]. Both methods allow for the removal of PCR in the preparation steps, not only rapidly decreasing DNA preparation time but also removing the risk of bias and error introduced via PCR (LIN 2012). Although the two methods differ on the signal used to detect nucleotide incorporation (SMRT uses fluorescence whereas nanopore use electric current), both methods collect the signals in real-time [10]. The average read length produced by SMRT is 1300bp and the potential is there for nanopore technologies to reach lengths of >5kbp and speeds of 1bp/ns [107].

There are also fourth-generation sequencing techniques, which are still in the experimental stage but are aimed at producing contextual sequencing, zooming in on individual transcripts within a cell or specific tissue [9]. These ideas are very much at a proof of concept phase, one such project referred to as "in *Situ*" involved the detection and genotyping of individual mRNA molecules in human and mouse cells resulting in the detection of a somatic point mutation and differentiation between members of a gene family [108].As human research forces these technological advancements, the plant science community will reap the benefit, making sequencing of large genome non-model plants more affordable and increasing the ability for *de novo* assembly.

1.5.6 Transcriptomics in the study of the biosynthesis of plant alkaloids and other high value compounds

With the rapid developments in sequencing technology, large collaborative efforts are now being made to study plants with secondary metabolites of biotechnological importance [2,6,109,110]. The objective is to discover metabolic pathways, compare genes in known gene families implicated in secondary metabolism and discover key genes/enzymes for production of valuable compounds.

Alkaloids is the term given to a diverse group of compounds containing nitrogen in a heterocyclic ring. Several key types of alkaloids have been used medicinally for centuries including benzylisoquinoline, monoterpene indole and amaryllidaceae alkaloids [1,8,110]. This includes the Monoterpene indole alkaloids (MIAs) vinblastine and vincristine (*Catharanthus roseus*) with anti-cancer properties, the BIA opiates including codeine (poppy) and the tropine and purine alkaloids nicotine and caffeine from tobacco (*Nicotiana tabacum*) and coffee (*Coffea arabica*) respectively [32] (see fig 1.6).

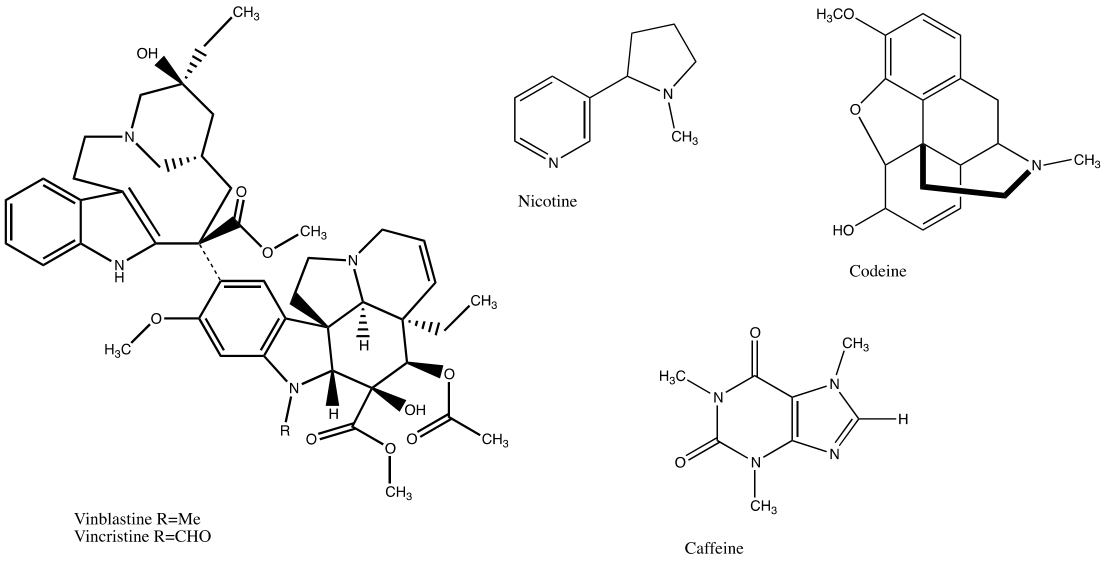


Figure 1-6 Chemical structure of widely used alkaloids (adapted from Takos *et al.,* 2013).

The compounds themselves are often species-specific but there is evidence that similar gene families are involved in the production of several compounds [5,111]. The opium poppy is one of the most widely studied alkaloid producing plants. A study in 2010 by Desgagne-Penix and colleagues produced a transcriptome using 454 sequencing that identified 427,369 ESTs (average length 462bp). The resulting assembly produced over 90,000 transcripts that were annotated using a BLAST database pipeline against well known databases such as UNIPROT [112]. This, along with proteomic data, resulted in the identification of genes involved in secondary metabolism [112]. Work specifically aimed at secondary metabolites has also been carried out in smaller projects in other non-model plants. For example, the work carried out by Guo *et al* (2013) on the traditional Chinese medicine plant *Dendrobium officinale* employed the 454 GS FLX titanium platform to produce 553,054 ESTs with an average length of 417bp. These were assembled into 36,907 unique contigs with 69.7% being annotated. The work carried out on this plant led to the annotation of 69 unique sequences relating to 25 genes in alkaloid backbone biosynthesis as seen in the KEGG database [5]. This study also reaffirmed the fact that key enzyme classes needed in the production of most secondary metabolites are cytochrome P450s, aminotransferases and methyltransferases and suggested that these may be co-expressed [5]. These will be discussed further in Chapter four.

One of the biggest collaborative transcriptomic plant projects is the PhytoMetaSyn collaboration [2,6] that has brought together thirteen research groups from seven Canadian institutes to produce genomic/metabolic resources for 75 plants known to produce high-value compounds. The project has discovered genes involved in biosynthesis of codeine and morphine, enzymes in diterpenoid biosynthesis for fragrances, undertaken *de novo* synthesis of sesquiterpenoids in yeast and characterized an enzyme involved in the biosynthesis of the bioactive compound thapsigargin [113-117].

A further large-scale project on 14 medicinally important plants has included investigation into the biosynthesis of MIAs by studying transcriptomes of pooled samples from different tissues of three MIAs producing plants, *Camtotheca acuminate, Catharanthus roseus and Rauvolfia serpentina* [110].

1.5.7 Transcript expression analysis in the search for agronomically and medicinally important compounds

Secondary metabolism, like primary metabolism, is a highly regulated process, with the most important mechanism in regulation suggested to be the amount of mRNA present [118]. Therefore, by looking at the differing levels of mRNA between individuals, or at different time points alongside metabolite profiling, it may be possible to predict individual genes involved in target pathways [119]. This can be particularly valuable in the study of alkaloid biosynthesis as alkaloids are often produced in specific tissue or at certain time points in development or under certain conditions.

Transcript level differences have been used in several projects for this purpose. One such study was carried out by Fridman *et al* in wild varieties of tomato. Differential expression studies on two varieties with clear differences in levels of methylketones led to the discovery of methylketone synthase 1. The EST was shown to be highly expressed in the variety PI126449 known to produce methylketones compared to LA1777 that did not produce the compounds [120].

As the gene families linked to secondary metabolism can contain numerous members that often use similar substrates and produce similar products it is difficult to determine which carry out specific reactions in alkaloid production [119]. By carrying out correlation studies looking at both mRNA expression and metabolite levels it is possible to predict links. This has been success in the study of BIA biosynthesis identifying and validating four NMTs in three species[121].

By comparing transcript level differences between Carlton and Andrew's Choice it may be possible to predict genes involved in amaryllidaceae alkaloid biosynthesis in this way.

1.5.8 SNP analysis in the search for high value compounds

As already discussed, the rapid development of DNA and specifically functional markers such as SNPs has revolutionized plant research and molecular breeding programs. Markers were originally used for genetic mapping but are now used in a variety of studies including characterizing germplasm, gene isolation, marker-assisted breeding and interrogation of target alleles [23]. DNA markers are composed of small regions of DNA that show polymorphism between individuals of the same species. They can be random (RDMs) phenotypic neutral markers or functional markers that show polymorphism in coding regions that affects the phenotype [23]. RDMs have been used in biodiversity studies such as the 1001 genome project in *Arabidopsis* but recombination can break the link between RDMs and target loci on alleles and so are limited in their diagnostic uses [122]. A more suitable method for representation of genetic variation is the use of functional markers as they are developed from coding regions that affect the phenotype. SNPs are the most abundant marker system, predicted to be an order of magnitude higher than that of SSRs, with a predicted average frequency of 1 SNP per 100-300bp [22]. Plant SNP studies have shown homologous SNPs to be more frequent. In wheat the frequency is about 1 per 20bp and 1 per 70bp in maize [22]. SNPs have been used in numerous plant studies to predict genes of agronomical value including the discovery of a SNP next to a waxy gene involved in amylose production in rice, a semi-dwarfing gene in rice as well as in a marker assisted breeding study into soybean resistance to cyst nematode [123-126]. It is hoped that similar SNP discovery methods can be used in daffodils to look for putative genes in alkaloid biosynthesis.

1.6 Conclusion and Hypothesis

1.6.1 Conclusion

Alkaloids are a large group of compounds with diverse medicinal uses and research into their biosynthesis is imperative to allow for further exploitation of their varied bioactivities. The use of RNA-Seq allows for a project of this scope. With the huge increase in transcriptomic data providing high sequence depth of coverage it is hoped that novel biosynthetic genes (and rare transcripts) for the unique compounds specific to alkaloid production in daffodils can be discovered [2].

1.6.2 Hypothesis

The biosynthesis of galanthamine and other alkaloids depends on controlled expression of genes for proteins that catalyse synthesis and sequestration of the alkaloids. Candidate genes can be identified following assembly and annotation of a *de novo* daffodil transcriptome. Annotation will rely on information in public databases from relatively distantly related species, including plants that also synthesise alkaloids. However, genes for secondary metabolism are known to fall into multi-membered families so additional information will be required to support candidates within gene families. A comparison will therefore also be made between two daffodil varieties that differ in galanthamine content, focusing on differences in transcript levels and identification of SNPs in homologs of genes that are predicted to be involved in alkaloid metabolism in other plants with the hypothesis that these differences may underlie the differences in secondary metabolites.

2 Chapter two - Construction of a reference transcriptome

2.1 Introduction

2.1.1 The use of Second-generation sequencing for de-novo assembly of transcriptomes in plants that produce medicinally important metabolites

One of the fundamental aspects of looking at the transcriptomes of plants of this nature is the need for *de novo* assembly of second-generation data due to the lack of genomic references. One of the most widely used technologies in 2010 to generate data for *de-novo* assembly was Roche 454 GS FLX titanium pyrosequencing. This method was advantageous since it produced longer reads than several of the other possible platforms making assembly more accurate [5] Several research groups have used 454 sequencing to produce a reference that can then be used alongside shorter reads from platforms that produce great number of reads to identify mutations and gene expression differences [2,12,127].

2.1.2 454 Sequencing technology

The 454 sequencing technology is explained in section 1.5.1.1. The resulting reads can then be used to create reference transcripts *via de* novo assembly.

2.1.3 Assemblers for *de novo* assembly of second-generation transcriptome data

There are two main strategies for assembling sequence data of this type, depending on whether the data is first mapped or assembled *de novo*. Mapping involves matching reads to a reference genome or transcriptome then merging overlapping reads into transcripts. However, for projects where no reference genome or closely related species is available (as for many plants at present) the sequence must be assembled using *de novo* methods. In theory *de novo* assembly should result in the complete reconstruction of the transcriptome allowing for the identification of all expressed genes, separate isoforms and expression levels. However, there are two main issues to be addressed when assembling transcript data. First, the second-generation techniques result in large numbers of short reads (20-100 million reads per sample in plants and animals) that increase assembly difficulty [128]. There is a high computational and memory cost involved in the assembly of these short reads, further increased with paired-end alignment and assemblies that can involve TBs of data (input and intermediate). Secondly, alternative splicing can result in shared exons between genes resulting in misassembled and concatenated sequences [128]. The 454 platform has its own mapping and assembly program known as GS *de novo* assembler more commonly called "Newbler" this program utilizes an modified version of overlap assembly.

2.1.4 Overlap assembly methods

This method was introduced with shotgun sequencing during the 1990s and has limited use with second-generation sequencing techniques due to the high computation costs of pairwise alignment on short reads and the overall increase of reads produced by the new methods [83]. Alongside the initial release of second-generation techniques the overlap method was altered to incorporate a clustering step that would allow it to be used with the higher number of long reads seen with 454 data [83]. The original process involved the creation of nodes, where each read represents a node, and edges are formed between two nodes when they overlap, followed by simplification steps to confirm overlap across both read orientations to remove transitive (false or redundant) nodes and edges resulting in a chain of nodes or "contig" [128]. As read number increased the computation time became too great to compare every read and so the reads were clustered into similar groups and overlap looked for within the clusters, current programs that use this improved overlap method include Mira, Phusion and Newbler as well as the Sanger assembler CAP3 [85,129-131].

2.1.5 Analysing transcriptome *de novo* assemblies

There are several key statistics that can be used to assess the assembly of *de novo* assemblies, these are explained briefly below.

Number of contigs assembled: This can only be compared when the number of contigs is known either through the use of simulated data or a genome reference. This is further complicated as contigs are often fragments of transcripts and so without a full reference this is not useful as an analysis of transcriptome assemblies.

N50: This is a measurement related to contig length; it is defined as the size of the smallest contig so that 50% of the total length of all contigs is represented in the contigs of size N50 or above. Although this has some use in transcriptomics, it is more insightful in genomic assembly as a higher N50 suggests less breaks in the genome, whereas transcriptomes are necessarily fragmented [128].

Reads mapped back to transcripts (RMBT): This is the number of raw or filtered reads that map back to the assembled transcripts, often given as a percentage [132].

Contigs mapped back to reference transcriptome or genome (CMBR): This is similar to RMBT but involves the mapping of the assembled contigs back to a reference transcriptome or genome. This is not possible in *de novo* assemblies without a reference but can be used with simulated data or known references to test the accuracy of the assemblers.

2.1.6 Annotation of *de novo* assemblies – The use of BLAST searches against known databases

Without an annotated genome in daffodils or closely related species it is important that the assembled transcripts are annotated. A method used by several projects is the use of BLAST to align transcripts to sequences from publically available databases. The use of databases such as UniProt, SwissProt, RefSeq, Interpro and Rfam have been used to produce annotations in plants such as olive, chili pepper and *Dendrobium officinale* with percentage of transcripts annotated ranging between 47 and 69%[2,5,14,101,133]. It is hoped that a similar pathway involving UniProt, TAIR, Rfam and RefSeq can be used to annotate the daffodil transcripts.

2.1.7 RNA extraction from plants with high phenol, sugar and secondary metabolite levels

The initial step in sequencing a transcriptome is the extraction of RNA for the production of a cDNA library. RNA extraction from plants is made difficult due to the varying levels and varieties of storage compounds and secondary products contained in plants [134]. There is currently no standard method of isolation that works for all plant species. Plant tissues that have high levels of polysaccharides, polyphenols and lipids have proven to be problematic when attempting to extract RNA with the widely used guanidium-phenol-chloroform extraction method [135]. Chang and colleagues addressed fundamental issues in RNA extraction using these methods and proposed an alternative method for the extraction of RNA from pine trees in 1993 [136]. One of the main problems associated with RNA extraction from many plants is the oxidation of the phenolic compounds that then bind irreversibly to nucleic acids and so precipitate with or degrade the RNA [135,136]. To combat this issue the use of PVP (polyvinylpyrrolidone), a strong polyphenolic compound binding agent that reduces nucleic acid degradation and β-mercaptoethanol as a reducing agent were introduced [135]. The methods also utilize CTAB (hexadecyltrimethylammonium bromide) instead of phenol to remove proteins thus minimizing the damage to poly A (+)-RNA that can occur in phenol extractions [136]. This CTAB method has since been modified by numerous groups to isolate RNA successfully from difficult plants such as bilberry (*Vaccinium myrtillus* L.) and peanuts (*Arachis hypogaea* L.) [134,135]. One additional modification by Dang and Chen was the introduction of a lithium chloride precipitation step that forms an RNA pellet leaving any DNA in the supernatant [135]. RNA extracted via these modified methods has proven to be suitable for cDNA preparation and RT-PCR reactions. Since daffodils have high levels of such problematic compounds a modified version of the 1993 method was used to extract RNA [134,135].

2.1.8 cDNA library preparation

A transcriptome consists of sequence data derived from mRNA, resulting in the analysis of only transcribed genes as previously discussed. The total RNA within the cell consists of >80% rRNA and, so that its three molecules do not dominate, this must be removed, or the mRNA fraction isolated, before cDNA can be prepared. Two recommended methods to isolate mRNA are mRNA selection and rRNA depletion. The poly (A) tail of most mRNAs can be used to select for mRNA by utilizing their interaction with poly(T) oligomers. This property is exploited in kits, such as the Invitrogen Dynabeads® mRNA purification kit, where the poly(T) is covalently bound to magnetic beads. Other RNA molecules, which lack the poly (A) tail will not bind to the beads and wash away (life-technologies, 2008). The rRNA depletion strategy is independent of polyadenylation or presence of a 5'-cap structure on the RNA and so offers a fuller isolation of the transcriptome [137]. The depletion method in Invitrogen's RiboMinus ™ plant kit for RNA-Seq removes rRNAs derived from cytoplasm (25/26S and 17/18S), chloroplast (23S and 16S), and mitochondrion (18S) from the total RNA. Both methods show high efficiency in the preparation of mRNA and therefore for a project of this type it would be beneficial to test both methods since neither method had been used with the Carlton variety previously.

2.2 Aims and Objectives

2.2.1 Overview of aims and objectives

Currently there is limited genomic or transcriptomic data available for daffodils, with no reference genome for daffodils or for any closely related plant. This project aimed to create an annotated transcriptome. In order to do this, a suitable method of RNA extraction, a modified CTAB method, was used to avoid contamination from oxidization of phenols. A cDNA library was then created for 454 pyrosequencing. This used two different methods of mRNA purification to determine which produced the greatest depletion in rRNA.

2.2.2 Data assembly and annotation

The sequenced library was assembled de-novo using the Newbler software supplied with the Roche/454 platform, which has been specifically developed to deal with 454 sequencing data. Finally, a BLAST database pipeline was created to annotate the assembly against several key public databases, UNIPROT, TAIR, RefSeq and Rfam. Several groups working on non-model plants, since it is possible to infer annotations from well-known plants from public databases have used this strategy. As discussed in section 2.1.6 strategies likes this have resulted in transcriptomes with over 60% of the contigs being annotated [5,111]. The annotated assembly has then been used for downstream analysis as a reference for read mapping, SNP discovery and gene expression profiling.

2.3 Methods

2.3.1 Plant material for reference transcriptome

Bulbs of *Narcissus pseudonarcissus* L. var. Carlton were obtained from Alzeim Ltd. The supplier to Alzeim was New Generation Daffodils (http://www.newgenerationdaffodils.com). Thirty bulbs of Carlton were planted in pots (36 cm diameter, 27.5 cm depth), 15 bulbs per pot on 14/12/2010. The bulbs were grown in a mixture of John Innes number 3 soil (Keith Singleton Horticulture, Cumbria, UK) and 3.0-6.0 mm Perlite supercoarse (William Sinclair horticulture Ltd, Lincoln.). Bulbs were planted at a depth of 15.2 cm. The pots were placed on the Institute roof thus experiencing normal weather conditions and inspected regularly, watering as necessary.

2.3.2 Basal plate extraction

Whole plants were dug up in mid April 2011 after the foliage had died back as this is the time predicted to show the highest level of galanthamine in the bulbs (personal communication with Dr X Chang). Four plants were washed in cold water and any remaining foliage removed. The roots were then removed and the basal plate cut from the bulb. This was then cut into small pieces of 1-2 mm thick, frozen in liquid nitrogen and stored at -80 °C until required. The basal plate was used in order to avoid the high levels of chloroplast transcripts expected in the bulb tissue causing rare or lowly expressed transcripts to be missed in the resulting sequencing.

2.3.3 Production of cDNA library

2.3.3.1 Total RNA extraction – CTAB method

The method was modified from Chang et al, (1993). The frozen basal plate tissue was ground in a pestle and mortar under liquid nitrogen. Extraction buffer (2% CTAB, 2% PVP40, 100mM Tris-HCl, 25mM EDTA, 2M NaCl, 0.5g l^{-1}, 2% β-mercaptoethanol) was warmed to 65°C and 2g powdered frozen tissue added before vortexing until homogeneous. The solutions were then incubated for 20 min at 65°C then placed on ice. The solutions were extracted 3 times in 10ml chloroform:isoamylalcohol (24:1). The phases were separated via centrifugation at 3696 *g* (Sorvall Legend RT with round buckets 75006533) at 4°C for 15 min carrying the supernatant (top layer) through between each extraction. Then 2.5 ml 10M lithium chloride was added to the supernatant and vortexed. The nucleic acid precipitate was allowed to form overnight at 4°C followed by centrifugation at 3696 *g* (Sorvall Legend RT with round buckets 75006533) for 30 min at 4°C. At this stage the RNA was purified using the RNeasy Plant mini kit (Qiagen) following the manufacturers RNA clean up protocol and dissolved in RNAse free water.

To determine the quality and quantity of RNA, 1μl was analyzed spectrophotometrically (Labtech NanoDrop® ND-1000 spectrophotometer). This determined concentration and OD ratios 260:280 and 260:230. The RNA was then analysed for degradation using formaldehyde gels following the method of Chang et al., (1993).

2.3.3.2 mRNA isolation

Two different protocols were used to isolate mRNA from the total RNA preparation to determine whether mRNA depletion or rRNA selection produced the higher yield and quality of mRNA for cDNA library preparation.

2.3.3.3 rRNA depletion

Sample 1 (concentration 10 μg total RNA in 1.1 μl) was rRNA depleted using the Ribominus kit (Invitrogen) according to the manufacturer's protocol. The method works by hybridizing rRNA molecules to locked nucleic acid probes for known rRNA molecules [138].

2.3.3.4 mRNA selection

Sample 2 (45 μg total RNA in 5 μl) was mRNA selected using the Dynabeads mRNA purification kit (Invitrogen) according to manufacturers protocol. This method involves the pairing of the poly A chains on the 3' end of mRNA to the oligo $(dT)_{25}$ residues on the surface of the beads [137].

2.3.3.5 cDNA Library preparation

Following rRNA depletion (sample 1) or mRNA selection (sample 2), both were used to construct cDNA libraries following the manufacturer's recommendations for the GS FLX titanium series cDNA Rapid Library Preparation Method (Roche). Both preparations started with 200 ng RNA. Since the samples would be pooled for sequencing, adaptor ligation was carried out to barcode the separate samples. Sample 1 was ligated to RL MID6, ATATCGCGAG, (Roche) and sample two to RL MID7, CGTGTCTCTA, (Roche).

The libraries were analyzed to check that the rRNA had been removed and that the fragmentation had been successful after depletion, fragmentation and at the end of the preparation using Agilent RNA 6000 Nano Kits. The resulting libraries were stored at -80°C until sequencing.

2.3.4 Creation of a reference transcriptome

2.3.4.1 454 Pyrosequencing

Both libraries were sequenced at the Centre for Genomic Research (CGR) at the University of Liverpool. This was carried out on the Roche 454 Pyrosequencing Titanium FLX series instrument. The libraries were amplified independently then pooled on half a plate for sequencing.

2.3.4.2 Assembly of reads from sequencing of barcoded libraries

The resulting sff files from the 454-pyrosequencing runs were assembled using the GS De Novo Assembler Newbler program (version 2.5). The program was run using the default settings. The two sff files corresponding to the barcoded libraries were first assembled independently and then assembled together.

2.3.4.3 Re-sequencing of sample 2 and full assembly from all GS FLX data

Sample 2 was re-sequenced at the CGR to give a greater coverage of the transcriptome due to the low read number of the original sequencing run.

A further assembly was carried out using the GS De Novo Assembler Newbler program (version 2.5). This assembly will be known as the full assembly from here on.

2.3.4.4 Manual annotation and the creation of an automated annotation pipeline for the joint and full assemblies using "full_annotation.pl"

The two assemblies were annotated separately using the same methodology as described below. The annotation was implemented using a variety of custom-built perl scripts and command line. The steps involved are shown in figure 2.1. The individual steps in the pipeline were combined to produce a single script that ran the whole pipeline from the command line in one step.

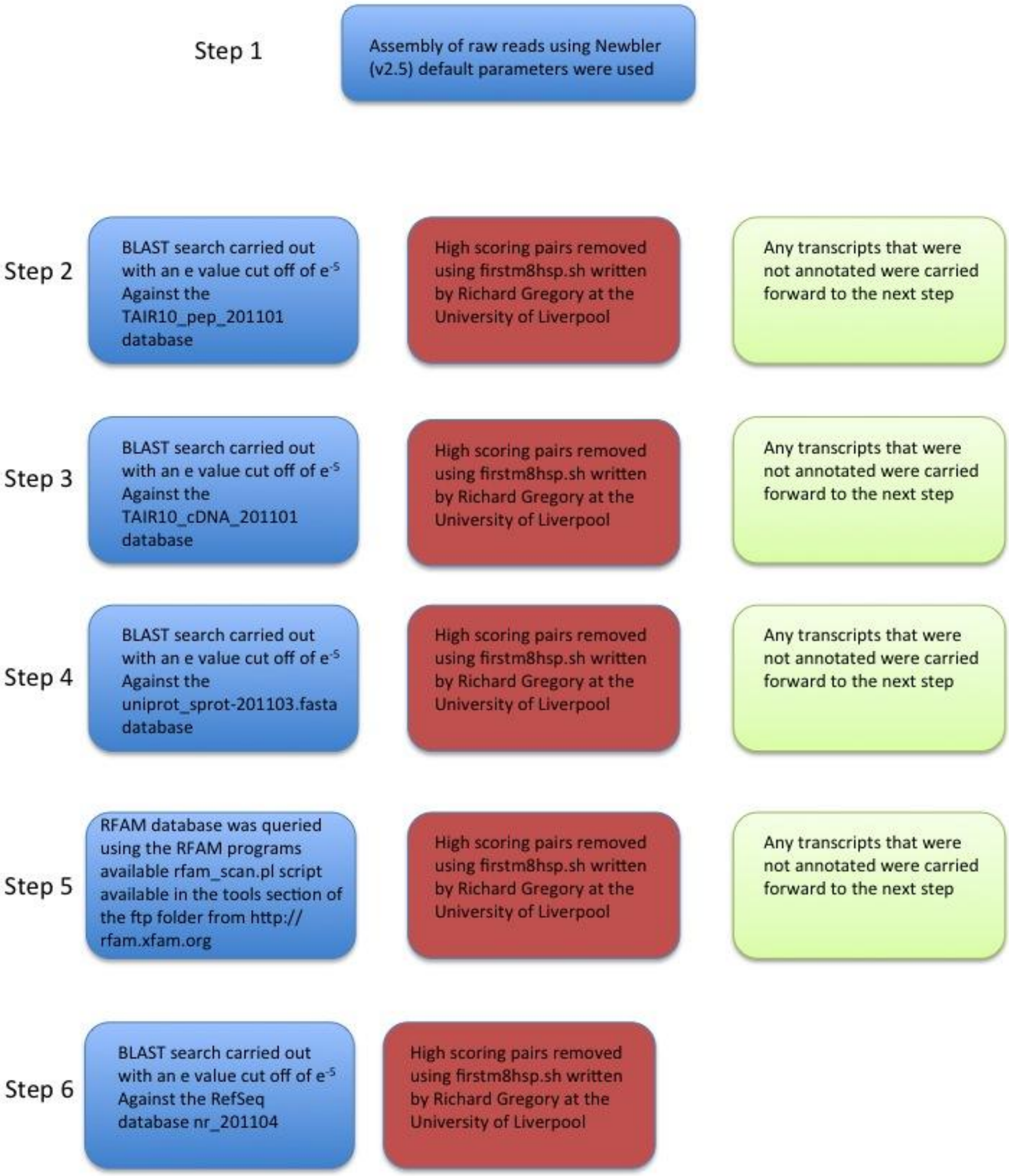

Step 1
Assembly of raw reads using Newbler (v2.5) default parameters were used
Step 2
BLAST search carried out with an e value cut off of e-5 Against the TAIR10_pep_201101 database
High scoring pairs removed using firstm8hsp.sh written by Richard Gregory at the University of Liverpool
Any transcripts that were not annotated were carried forward to the next step
Step 3
BLAST search carried out with an e value cut off of e-5 Against the TAIR10_cDNA_201101 database
High scoring pairs removed using firstm8hsp.sh written by Richard Gregory at the University of Liverpool
Any transcripts that were not annotated were carried forward to the next step
Step 4
BLAST search carried out with an e value cut off of e-5 Against the uniprot_sprot-201103.fasta database
High scoring pairs removed using firstm8hsp.sh written by Richard Gregory at the University of Liverpool
Any transcripts that were not annotated were carried forward to the next step
Step 5
RFAM database was queried using the RFAM programs available rfam_scan.pl script available in the tools section of the ftp folder from http://rfam.xfam.org
High scoring pairs removed using firstm8hsp.sh written by Richard Gregory at the University of Liverpool
Any transcripts that were not annotated were carried forward to the next step
Step 6
BLAST search carried out with an e value cut off of e-5 Against the RefSeq database nr_201104
High scoring pairs removed using firstm8hsp.sh written by Richard Gregory at the University of Liverpool

Figure 2-1 Annotation pipeline using "full_annotation.pl".

The steps were implemented using perl scripts and command line prompts. The eventual aim would be to have this as a simple pipeline that can be used by non-bioinformaticians for future analysis. The script “full_annotation.pl” can be seen on appendix disc. The steps in blue were carried out using the command line version of BLAST 2.2.27+, using BLASTX for steps 2,4 and 6 and BLASTN for step 3 with the –m 8 tabulated output option and –b 1 and –v 1 options for upper limits on number of database sequences to show alignments for and number of one-line descriptions to show. The steps in red were carried out via a perl script created by Richard Gregory of the CGR at the University of Liverpool that removes high scoring pair results so that each transcript has only one hit. The steps in green were implemented using a variety of UNIX commands and perl hashes to pull out the sequences from the fasta file that had no hits.

2.4 Results

2.4.1 Creation of a cDNA library

2.4.1.1 RNA extraction – CTAB Method

Analysis for total RNA quality after extraction and purification is shown in Figures 2.2 and 2.3.

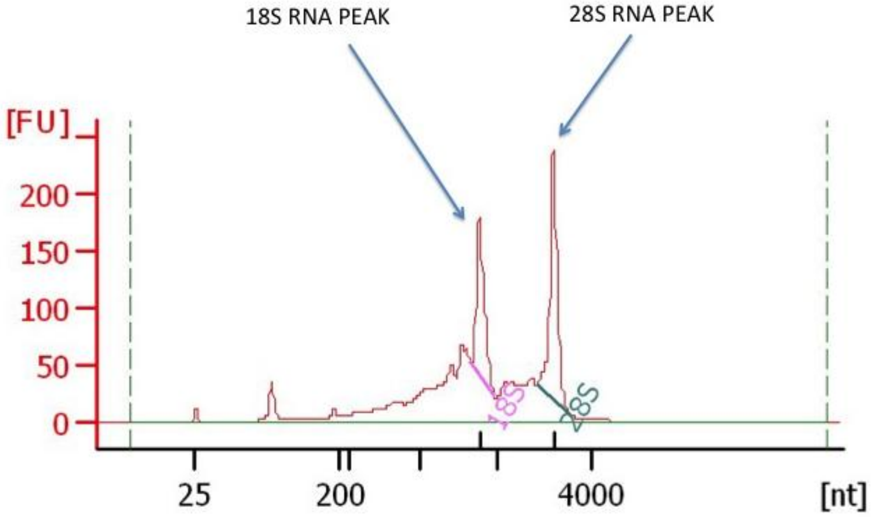


Figure 2-2 Total RNA electropherogram for sample 1

The rRNA ratio (28S/18S) was 1.4 with an RNA integrity number (RIN) of 7.3 and a final concentration of 9.25 µg µl^{-1}. FU = fluorescence units.

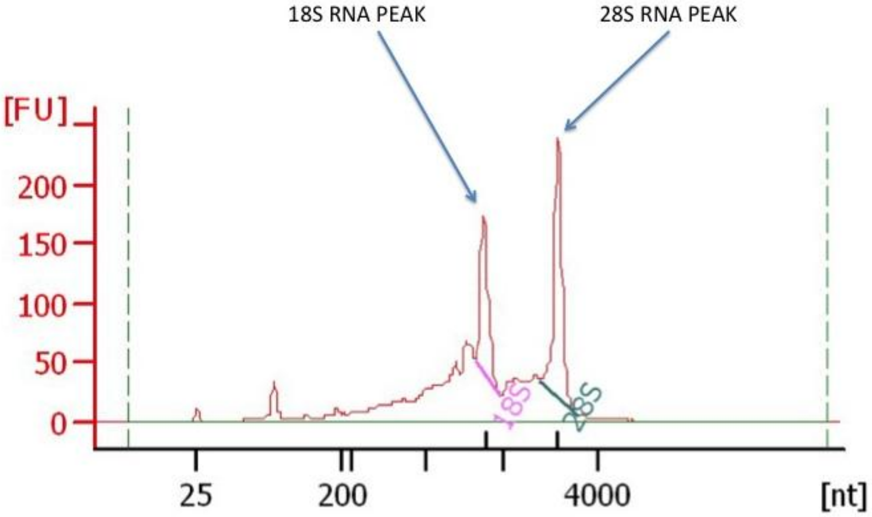


Figure 2-3 Total RNA electropherogram for sample 2

The rRNA ratio (28S/18S) was 1.5 with an RNA integrity number of 7.3 and a final concentration of 9.30 µg µl^{-1}.

The RNA integrity (RIN) number was above the degradation threshold of 6. At this point the 28S/18S ratio remains close to 2 and the baseline signal is relatively low. The 28S peak is clearly more intense than the 18S peak (ideal theoretical ratio is 2:1 based on the observed ratio in mammalian RNA). The 18S and 28S are the small and large subunits of the ribosomal RNA. It is rare to see ratios of exactly 2 or higher due to the relative instability of the 28S RNA compared to the 18S RNA. Therefore both samples were deemed to be of a high enough yield and quality to continue with the mRNA isolation step prior to cDNA library preparation.

2.4.1.2 mRNA isolation

2.4.1.2.1 rRNA depletion of sample 1

Sample 1 was rRNA depleted using the Ribominus kit (Invitrogen). The method required a starting amount of 10 µg of total RNA. The depletion was carried out and the sample was then fragmented to the required range for sequencing (600-1200bp) at a final concentration of 40 ng µl^{-1}. The recovery at this step was 4% (optimum range 1-10%). As no 28S or 18S peaks could be seen on the Agilent gel and 28S/18S ratio was 0.0, (Figs 2.4, 2.5) it suggested over 95% rRNA depletion and the lack of obvious peaks suggested successful fragmentation.

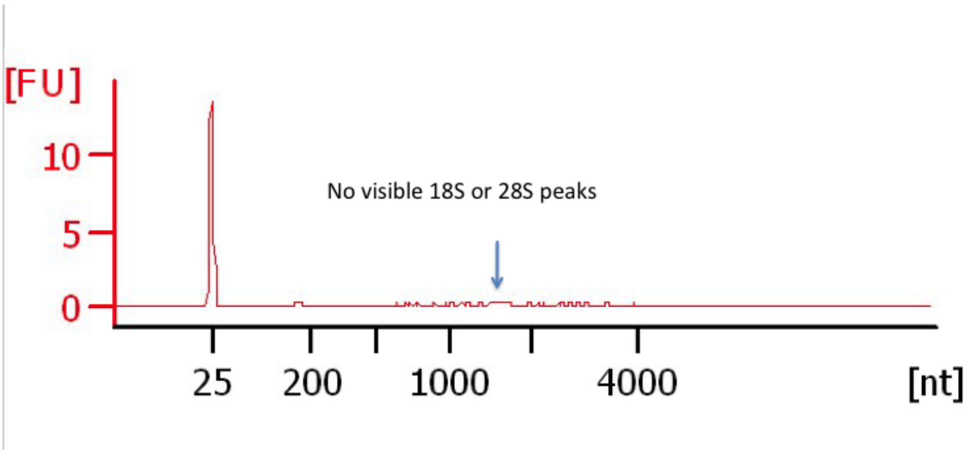


Figure 2-4 Depleted sample 1 electropherogram

The lack of obvious peaks except for the marker peak at 25nt suggests successful depletion.

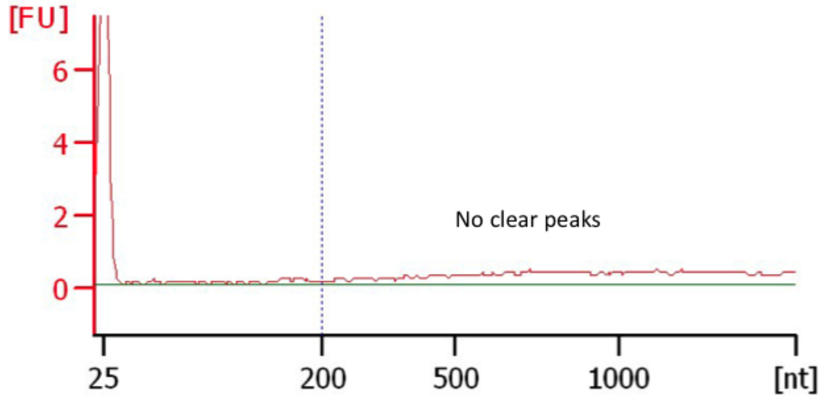


Figure 2-5 Fragmented sample 1 electropherogram

The lack of any obvious peaks suggests fragmentation was successful.

2.4.1.2.2 mRNA selection of sample 2

The second sample of total RNA was subjected to mRNA selection using the Dynabeads mRNA purification kit (Invitrogen). The starting amount was 45 µg total RNA in 5 µl. The resulting 28s:18s ratio of 0.7 suggested an 80% depletion rate with a recovery of 0.8% total RNA. This is below the optimum level of 2-4% but was considered acceptable at this stage due to the limited availability of the sample material.

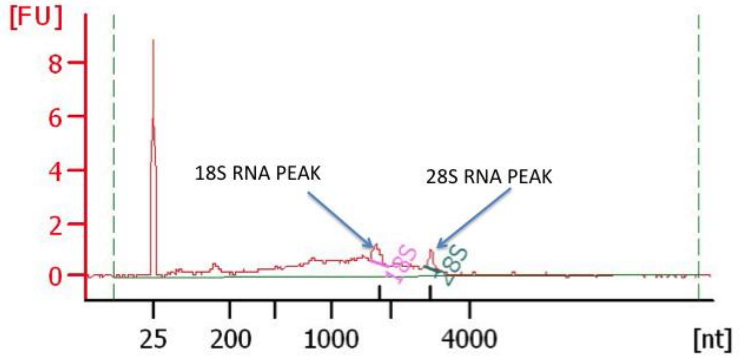


Figure 2-6 Fragmented sample 2 electropherogram

28s/18s ratio suggests good fragmentation with 80% depletion. The marker peak is clearly visible at 25nt. The small 18s and 28s peaks suggest less depletion than that seen for sample 1. This is reflected in the calculated depletion of 80%.

Both fragmented samples were used to produce cDNA libraries for sequencing.

2.4.1.3 cDNA library preparation

2.4.1.3.1 Sample 1

A fragmented cDNA library was constructed using the GS FLX titanium series cDNA Rapid Library Preparation protocol. The final concentration was 6 x 10^8. The resulting cDNA library fragments ranged between 500-2000 bp with an average size of 820 bp. This is well within the working range of the 454 platform.

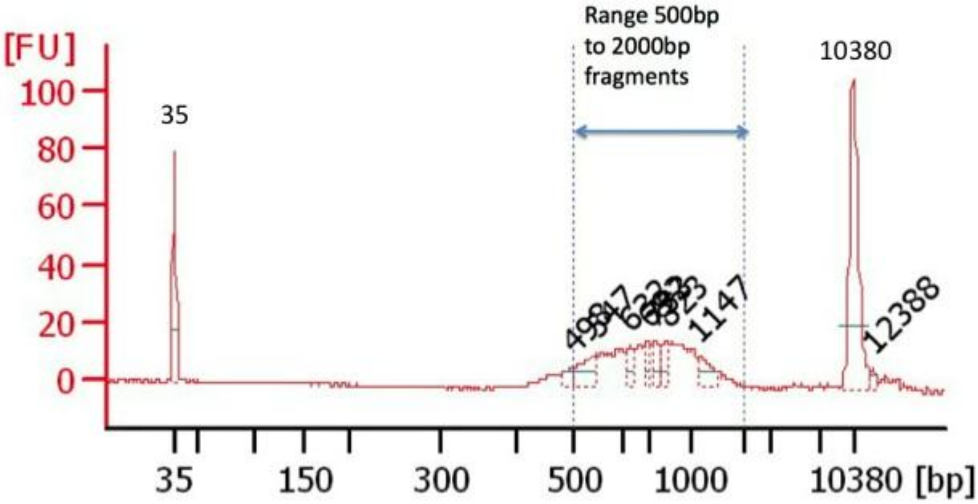


Figure 2-7 High sensitivity DNA assay electropherogram for cDNA library from sample 1

The two peaks at 35 and 10380 are marker peaks.

2.4.1.3.2 Sample 2

A cDNA library was then constructed in the same way as sample 1 with MID7 adaptors (CGTGTCTCTA). The final library concentration was 6.5 x 10^8. The fragments ranged from 476-1983 bp with an average of 825 bp, within the working range of 454 sequencing.

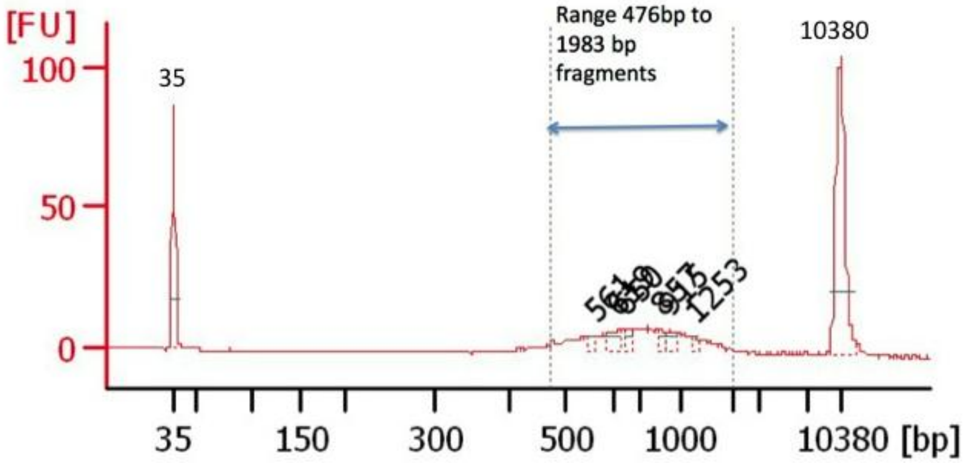


Figure 2-8 High sensitivity DNA assay electropherogram for cDNA library from sample 2

Library consisted of fragments in the range of 476-1983 bp with an average length of 825 bp. The two dominant peaks at 35 and 10380 are marker peaks.

2.4.2 Production of a reference transcriptome

2.4.2.1 Assembly of reads from sequencing of the barcoded libraries

The initial (joint) assembly of the 454 -pyrosequencing data of sample 1 and 2 via Newbler (v.2.5) resulted in 226,848 ESTs of which 189,297 were assembled. The assembled reads produced 32,853 transcripts consisting of 1728 (5%) contigs and 31,224 (95%) singletons. The average contig size was 733bp with an average trimmed raw read length of 397.858bp and the largest contig was 6849bp.

2.4.2.2 Mapping of barcoded libraries to the joint assembly

The two separately barcoded libraries were assembled independently and mapped back to the joint assembly. This was used to test the coverage of the transcriptome by each library. The results of assembly and mapping to joint assembly are shown in tables 2.1 and 2.2.

Table 2-1 Assembly comparisons of the two samples

	Sample 1	Sample 2
Total number of reads	150708	76190
Number of aligned reads to joint assembly	132064 (88%)	55290 (73%)
Number of assembled reads	116607 (77%)	49752 (65%)
Number of singletons	14465	18750
Number of contigs	463	1111
Number of reads too short to assemble	3760	1703
Average contig length	780bp	719bp
N50	714bp	700bp
Largest contig	6669bp	5328bp
Average Length of trimmed reads	405.1bp	383.389bp

Table 2-2 Comparison of the two samples mapping to joint assembly

	Sample 1	Sample 2
Total number of reads mapped to joint assembly	77980 (97%) Inferred read error = 0.96%	55087(98%) inferred read error = 1.06%
Number of contigs and singletons partially mapped	7653 (51%)	10606 (53%)
Percentage of reads fully mapped	21.49%	32.74%
Percentage of reads not mapped	0.02%	0.08%

From tables 2.1 and 2.2 it can be seen that both samples produced similar assemblies and mapping. The contigs were annotated using the UNIPROT retrieve program (http://www.uniprot.org/?tab=mapping&tab=batch). This was implemented to look for ribosomal hits to make sure that rRNA depletion had been successful and only mRNA had been sequenced. Neither method revealed any ribosomal RNA sequences suggesting that both methods of mRNA isolation were successful. Sample 2 was sent for further sequencing to increase coverage due to a sequencing error in the initial run leading to lower read numbers than seen for sample 1.

2.4.2.3 Re-sequencing of the sample 2

The re-sequencing of the sample resulted in 87763 reads with an average (trimmed) read length of 385bp. This shows only a small improvement on read number over the initial sequencing and could suggest an error in the cDNA library preparation. It was initially assembled alone to give 1082 contigs with an average length of 862bp. It was decided that a full assembly involving the trimmed reads from all three sequencing runs (Sample 1 and Sample 2 the Sample 2 repeat) should be carried out to build a consensus reference.

2.4.2.4 Assembly of the re-sequenced data with the initial barcoded data (full assembly)

The initial samples along with the re-sequenced Sample 2 data were assembled using Newbler (v2.6) that resulted in the assembly of 161,739 of the 314,591 ESTs produced. The assembled reads produced 41,302 (91%) singletons and 4022 (9%) contigs. The average contig size was 755bp with an average trimmed raw read length of 395bp and the largest contig was 6932bp.

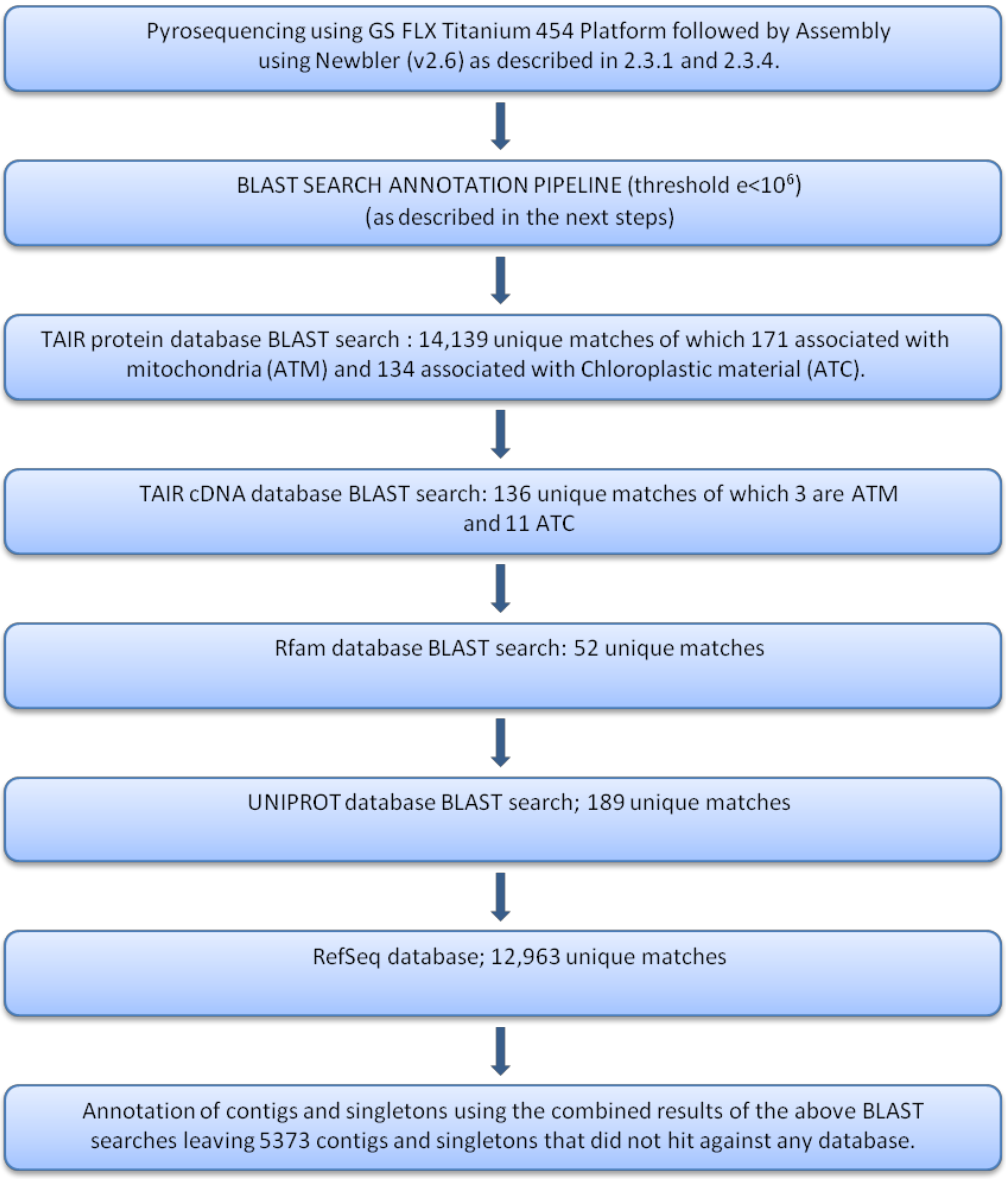
2.4.2.5 Mapping of thee sequencing samples to full assembly

Table 2-3 Mapping of the three separate sequencing samples to the full assembly.

	Sample 1	**Sample 2**	**Sample 2 repeat**
Total number of reads mapped to joint assembly	134676 (89.37%)	61028(80.15%)	66037 (75.26%))
Inferred read error (%)	0.52	0.72	1.23

2.4.3 Annotation and creation of an automated annotation pipeline for the joint and full assemblies

2.4.3.1 Joint assembly



Pyrosequencing using GS FLX Titanium 454 Platform followed by Assembly using Newbler (v2.6) as described in 2.3.1 and 2.3.4.

BLAST SEARCH ANNOTATION PIPELINE (threshold $e<10^{6}$) (as described in the next steps)

TAIR protein database BLAST search : 14,139 unique matches of which 171 associated with mitochondria (ATM) and 134 associated with Chloroplastic material (ATC).

TAIR cDNA database BLAST search: 136 unique matches of which 3 are ATM and 11 ATC

Rfam database BLAST search: 52 unique matches

UNIPROT database BLAST search; 189 unique matches

RefSeq database; 12,963 unique matches

Annotation of contigs and singletons using the combined results of the above BLAST searches leaving 5373 contigs and singletons that did not hit against any database.

Figure 2-9 Pipeline to show the steps of the annotation of the joint assembly.

The pipeline resulted in the annotation of 27479 of the 32852 singletons and contigs (84%).

Figure 2.9 shows the steps involved and the results of the annotation pipeline. The number of hits shown for each BLAST search is post clean up (removal of high scoring pairs and low scoring results). The further investigation of this annotation into functionality and possible putative genes involved in secondary metabolism is described in chapter four.

2.4.3.2 *Full assembly*

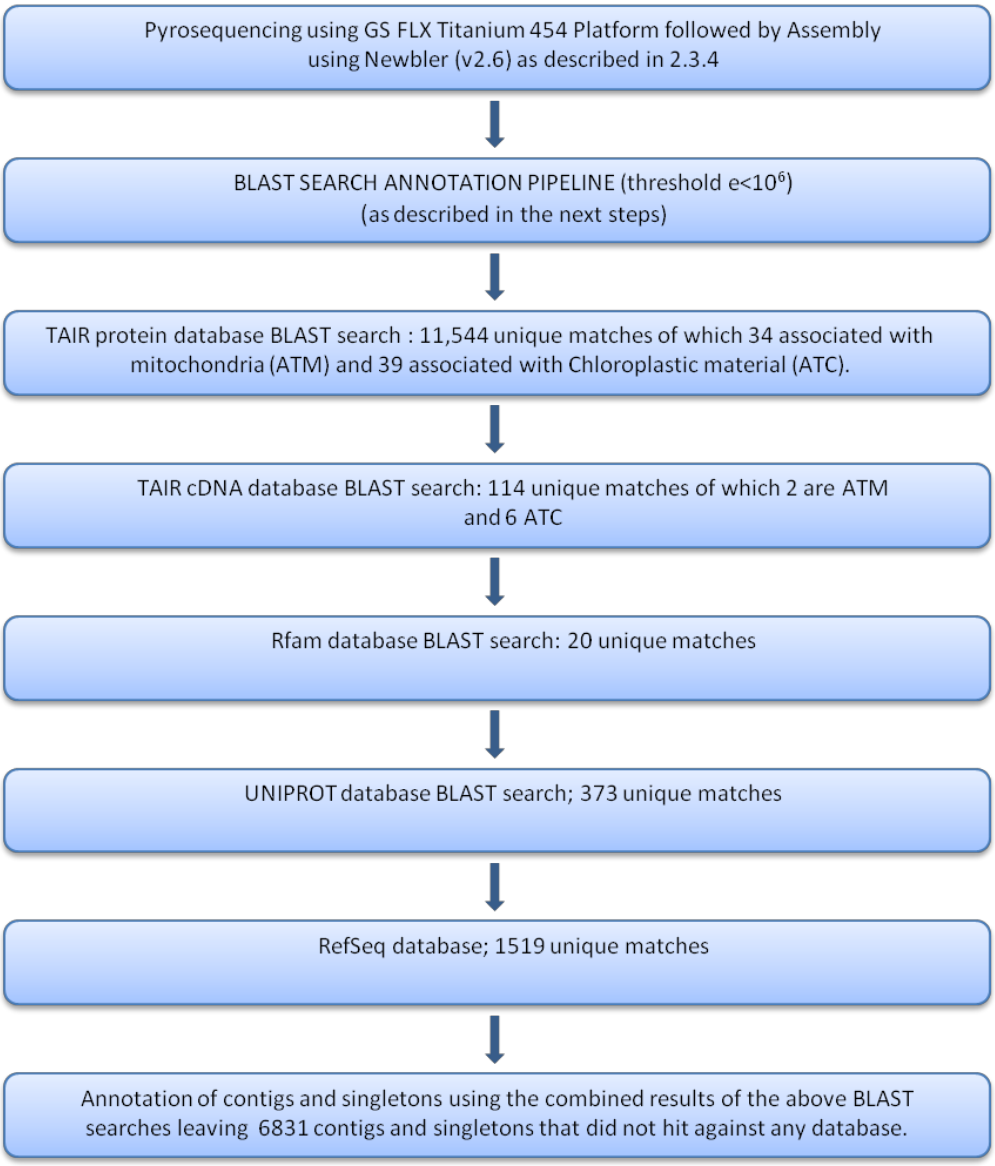

Pyrosequencing using GS FLX Titanium 454 Platform followed by Assembly using Newbler (v2.6) as described in 2.3.4
BLAST SEARCH ANNOTATION PIPELINE (threshold $e<10^{6}$) (as described in the next steps)
TAIR protein database BLAST search : 11,544 unique matches of which 34 associated with mitochondria (ATM) and 39 associated with Chloroplastic material (ATC).
TAIR cDNA database BLAST search: 114 unique matches of which 2 are ATM and 6 ATC
Rfam database BLAST search: 20 unique matches
UNIPROT database BLAST search; 373 unique matches
RefSeq database; 1519 unique matches
Annotation of contigs and singletons using the combined results of the above BLAST searches leaving 6831 contigs and singletons that did not hit against any database.

Figure 2-10 Pipeline to show the steps involved in the annotation of the full assembly.

The pipeline resulted in the annotation of 67% of the singletons and contigs assembled.

Figure 2.10 shows the steps involved and the results of the annotation pipeline. The number of hits shown for each BLAST search is post clean up (removal of high scoring pairs and low scoring results so that only the top hit for each contig is used for annotation). The next step is further investigation of the annotated hits (see Chapter four).

2.5 Discussion

RNA extraction from plants with high levels of phenols, sugars and secondary metabolites such as *N. pseudonarcissus* Carlton is a difficult process. The use of the CTAB method with the modified LiCl step [127] followed by RNA clean up produced high yields of RNA from this specific variety. This was tested using the Labtech NanoDrop® ND-1000 spectrophotometer and an Agilent RNA 6000 Nano Kit. The results were comparable to similar extractions in other plants in that sample 1 had an RIN of 7.3 and 28S/18S ratio of 1.4 with an A_{260}/A_{280} ratio of 1.97 whereas sample 2 had a RIN of 7.3, 28S/18S ratio of 1.5 and an A_{260}/A_{280} ratio of 1.98. These are similar to the ranges found in peanut (A_{260}/A_{280} 1.99-2.06) and soybean, sunflower, oil seed and canola (A_{260}/A_{280} 2.06-2.17) [126]. The ratios were also similar to the original methodology paper by Chang *et al* in 1993 on pine trees (A_{260}/A_{280} 1.7-2.0). The final concentrations of RNA of sample 1 (9.25μg μl^{-1}) and sample 2 (9.30 μg μl^{-1}) along with the A_{260}/A_{280} ratios were considered suitable for cDNA library preparation as the minimum amount of RNA required is 200ng with an OD (A_{260}/A_{280}) of ≥1.8 [139]. The method however, is not consistent, with variation in yield and RNA degradation level observed (not shown). Further evaluation of RNA extraction methods and kits would be beneficial and therefore further trials were carried out for RNA extraction for a second variety of *Narcissus pseudonarcissus* var. Andrew's Choice.

The two methods of mRNA isolation (rRNA depletion and mRNA selection) used standard methods and kits, and both gave acceptable levels of rRNA depletion, (95% and 80% respectively). The mRNA recovery step for sample one was within range (1-10%). The second sample was below the normally accepted level (2-4%) but after discussion with the sequencing team at the University of Liverpool Centre for Genomic Research it was decided that the sample was still viable and would be used to produce a sequencing library. This was in part due to the fact that it was of great importance and there was very little tissue available at the time from the basal plate of Carlton in the right growth phase. Two similarly sized cDNA libraries were finally produced (sample 1= 6x10^{8}, 500-2000bp with an average fragment length of 820bp; sample 2 = 6.5x10^{8}, 476-1983bp with an average fragment length of 820bp).

The reads resulting from the sequencing of the two samples were combined and assembled. The initial assembly resulted in 226898 ESTs of which 189,297 (average trimmed read length of ~398bp) were assembled into 32,853 transcripts with an average contig length of 733bp. The raw reads from the two samples were mapped back to the joint assembly and both resulted in over 97% RMBT which suggests a good assembly of raw reads [140]. However the majority of assembled transcripts were singletons and so this may suggest poor overlap between the two samples, as they are both from the same variety and the same tissue type you would expect some overlap of reads. The project would benefit from further sequencing or perhaps assembly of the separate samples using Newbler and a second assembly of the transcripts produced via a program such as Cap3 [131]. This would allow for the collapsing of singletons and contigs into bigger transcripts and has been used in other similar projects, it would also make repeats or chimeric sequences more apparent as the original references could be mapped to each other and any reads that do not map could then be used alongside the assembled transcripts in a second assembly [110].

Some additional sequencing was carried out on sample 2 as the CGR noted there was an error during sequencing that might explain the much lower number of raw reads produced for this sample. However the re-sequencing resulted in similar numbers of reads and a separate assembly of this run showed similar contig numbers (1082, average length 862bp) and average raw read lengths (385bp). The lower number of reads was therefore not attributable to sequencing error but was possibly caused by a poor cDNA library preparation as suggested by the lower than ideal mRNA recovery step of sample 2 (section 2.4.1.1.2).

The final full assembly encompassing all three sequencing sets (sample 1 and the two runs of sample 2 resulted in just under 100,000 reads of which 161,739 were assembled to give just over 45,000 transcripts. The mapping of the three samples showed that sample 2 mapped back to the reference better than the two sample 2 runs (89% compared with 80 and 70%) suggesting that more of the sample 1 reads were used in the assembly, again suggesting that sample 2 may not have provided a good quality cDNA library. Overall however, these results are comparable to those seen in similar projects such as that previously discussed in *Dendrobium officinale* where the 454 sequencing resulted in 553,054 reads assembled to 36,907 transcripts with an average length of 417 [5].

In order to gain any further understanding of this transcriptome, the transcripts must be annotated. No reference genome is available for *N. pseudonarcissus* or related species, so this transcriptome must be annotated *de novo* Several projects involved in the production of *de novo* assemblies of transcriptomes use widely available public databases to assign annotations to the novel transcriptomes. BLAST searches against TAIR, UNIPROT, RefSeq and Rfam were therefore carried out resulting in 67% of the transcripts being annotated; similar to other projects of this nature [5,100].

The reference created in this chapter (known as the 454 reference or assembly throughout later chapters) is a building block for the investigation into alkaloid biosynthesis in daffodils. In order to identify transcripts linked to alkaloid production, it is vital that this reference and its annotation are built upon. It is not only useful to annotation *de novo* assemblies via homology such as BLAST searches but also to confer functionality and compare to other alkaloid producing systems. This is carried out in the following chapters.

3 Chapter three: Implementing Second-generation sequencing data for discovering polymorphic and transcript level differences between two varieties of *Narcissus pseudonarcissus* with distinct differences in galanthamine levels.

3.1 Introduction

As discussed in Chapter One the development of second-generation sequencing techniques such as the Roche 454 pyrosequencing and Illumina platforms have allowed investigation of transcriptomes and genomes of non-model organisms with large genomes. With the influx of data from non-model plants, research is shifting towards specific areas of interest. One area, which is rapidly developing, is the use of second-generation sequencing to investigate the biosynthesis of medicinally important secondary metabolites [2].

This project initially employed 454-pyrosequencing to produce a reference transcriptome for the *Narcissus pseudonarcissus* variety Carlton, as described in Chapter Two. The resulting reference transcriptome could then be used to look for genes linked to galanthamine production.

3.1.1 Illumina Sequencing

Illumina, like 454 sequencing, is a “sequencing by synthesis” method. Sequencing templates are clustered on flow cell surfaces and fluorescently labeled dNTPS are added one at a time. Incorporation bias is avoided and raw error rates are reduced by having all 4 dNTPs in equal amounts, adding only one nucleotide per cycle and measuring fluorescence at the end of each cycle [80]. The desire to fully utilize the large amounts of data created by Illumina to capture the full transcript profile has led to the creation of *de novo* assemblers developed specifically to assemble these shorter reads.

3.1.2 Assembly of Illumina reads

With the introduction of Illumina sequencing technologies and the vastly increased read number (~160x that of 454 in 2010) the overlap method of assembly (section 2.1.4) became increasingly redundant and so a novel method using de Bruijn graphs was developed [25].

3.1.2.1 Assemblers that use de Bruijn graphs

For de Bruijn graphs, reads are broken down into nodes of a chosen length "k" (known as k-mers) connected by edges if nodes overlap by k-1 nucleotides. This results in all solutions by which linear sequence can be reconstructed and in the case of transcriptomes each path in the graph is a possible transcript [140]. A scoring algorithm can then be used with the original sequence and any mate pair information to remove nonsensical solutions [132]. The basis of these aligners is the identification of overlap between "k-mers" if there are differences the graphs branch, forming split ends if no further identity is seen or bubbles in the case of single nucleotide differences or INDELs [83]. One issue of these bubbles is the difficulty in distinguishing whether they are caused by natural variation of sequencing error. Further deviation from the linear graphs can occur due to alternative splicing or improper trimming or filtering of low quality reads, this makes aligning reads more difficult especially for transcriptomes [83]. Within genome assemblies issues such as sequencing error can be resolved by looking at the coverage of each base and removing k-mers with low coverage, however in transcriptomes the coverage is intrinsically uneven [83]. Coverage is altered by SNPs, alternative splicing and transcripts that have naturally low expression meaning real transcripts or SNPs can be lost if using a traditional coverage cut-off [83]. It is possible to use a normalized library for transcriptomes but quantification is lost and so for a project of this nature looking at gene expression cannot used such a strategy. In order to confront these issues, add-ons, software looking at quality per base prior to assembly to remove erroneous reads and novel assemblers specifically designed for *de novo* transcriptome assemblies have been introduced. Several comparisons have been carried out on the performance of these assemblers and the findings will be briefly described below. Trinity and SOAPdenovo-Trans will then be explained in further detail.

3.1.2.2 Comparison of assemblers

In the construction of de Bruijn graphs there are clear differences between genomes and transcriptomes. In genomes only a relatively small number of large connected sequence graphs are constructed to show read connections across the entire chromosome [132]. For transcriptions however, due to the complexity of non-overlapping loci caused by non-transcription of intergenic sequence, large numbers of individual graphs are generated thus greatly increasing computation time[132]. A balance is needed that allows for the removal of variation caused by sequencing error and other technical issues while retaining true biological variation. Along with memory usage and computational time the key statistics discussed in 2.1.5 are used to compare *de novo* transcriptome assemblers (number of contigs assembled, N50, RMBT and CMBR).

Several assemblers have been compared using these statistics; the most widely compared are Trinity, a program developed specifically for *de novo* transcriptome assembly, Oases (Velvet) originally developed to resolve pre-existing errors and repeats in short read genome assembly, SOAPdenovo, designed as part of the SOAP suite from the Beijing Genomics Institute for human genome *de novo* assembly, ABySS, designed to speed up short read assembly by assembling the reads in parallel, Scripture, designed for *ab initio* reconstruction of mammalian cell RNA genome and Cufflinks an open source algorithm designed to discover transcripts and estimate abundance in one [132,141-145]. Of these only Trinity was specifically designed for *de novo* transcriptome assembly although SOAPdenovo now has a transcriptome assembler available called SOAPdenovo-Trans. (This has not been extensively compared to the others and so is not discussed here but see section 3.1.2.4 for more details). Clarke and colleagues carried out comparisons of Trinity, ABySS, Velvet and its transcriptome algorithm Oases using simulated and real RNA-Seq data on artificial RNA templates and human transcripts [128]. The study found that no method was superior to all others. Trinity performed well across all tests but was not consistent in assembling full-length transcripts, although it had an N50 just above the average transcript length, 90% RMBT but only 46% CMBR [128]. Grabherr *et al* compared their assembler, Trinity, to ABySS (and its modified transcriptome assembler TRANS-ABySS), SOAPdenovo, Scripture and Cufflinks. The study showed that Trinity out-performed the other *de novo* assemblers and was comparable to genome assemblers when looking at fission yeast, mice and whitefly (*Bemisia tabaci*) [132]. It showed higher sensitivity (number of reference transcripts successfully reconstructed to full length) than the other assemblers tested and was comparable for both transcriptome coverage and RMBT rates [132]. Zhao *et al* also compared the assemblers Trinity, SOAPdenovo, Oases and ABySS. Trinity resulted in the best RMBT percentage and performed well across the tests but had a long run-time (Oases had the longest). The shortest run-time was that of SOAPdenovo but this performed poorly in other aspects. ABySS showed a balance between run-time and performance but Trinity was nevertheless considered the best single k-mer method [140].

Trinity was specifically designed for *de novo* assembly of transcriptomes and has performed well compared to other assemblers. It was therefore decided to assemble the *Narcissus* data using Trinity. However, there has been little assessment of the 2013 SOAPdenovo-Trans transcriptome assembler so this was compared with Trinity.

3.1.2.3 Trinity

Trinity uses de Bruijn graphs and an enumeration algorithm for scoring all possible paths, along with actual read and paired-end information to remove nonsensical edges to leave plausible transcripts or isoforms [140]. The use of paired-end read data allows for better resolution of miss-assemblies as it increases the distance that Trinity can look for ambiguities [132]. Trinity is made up of three modules; Inchworm assembles the reads into a unique set of transcripts employing a ‘greedy’ k-mer approach, only choosing the ‘best’ representation for a set of alternative variants that share k-mers (due to alternate splicing, gene duplication or allelic variation). Chrysalis examines the complexity of overlap between variants, which was ignored by Inchworm. It clusters related contigs and constructs a graph for each cluster. Then, Butterfly is the final step that analyses all the graphs, and paths within them, taking into consideration read-pairs. This step reports all possible transcripts, theoretically resolving alternative splicing isoforms, and those transcripts representing paralogous genes [132].

3.1.2.4 SOAPdenovo-Trans

This modified version of SOAPdenovo2 was created for use with the 1000 Plants project (www.onekp.com) and combines the de Bruijn graph method with a local error removal method similar to that from Trinity and the graph traversal method from Oases [146]. The graph traversal method is an algorithm that analyses the clustered sub-graphs after graph simplification to generate all possible transcripts from linear, fork and bubble paths [146]. In initial testing SOAPdenovo-Trans showed slightly higher but comparable performance to Trinity on small and large rice datasets with 91.8% and 89.5% correct transcript assembly (Trinity-84.6% and 81.5%) [146]. The assembler consists of two main modules. Firstly, Contig assembly uses the same de Bruijn graph technique as SOAPdenovo but carries out the sequence error removal step in two ways. Initially low-frequency k-mers and edges are removed using the same global removal step as SOAPdenovo [146]. However, transcriptomes intrinsically show varying levels of expression and if a sequencing error occurs in highly expressed genes this will be missed via the global threshold. Since these erroneous k-mers in highly expressed transcripts may be above the error threshold, they are removed locally as in Trinity. By using a non-constant threshold ≤ 5% of the total or maximum depth of adjacent graph elements it is possible to remove these highly expressed sequence errors and account for variable expression seen in transcriptomes.

Transcript assembly is carried out in four steps and incorporates scaffolding methods from SOAPdenovo and graph traversal from Oases. Scaffold construction uses paired-ends and maps reads back to the contigs to form linkages. Graph simplification then removes sequence errors and short repeated regions. There is also more stringent linearization of the transcripts to account for alternative splicing. Short contigs (≤ 100bp) are removed but this results in gaps that must be filled later. Graph traversal then clusters contigs using the Oases algorithm to predict possible transcripts. Finally the gap filling method from SOAPdenovo is carried out with paired-end information to create a consensus.

3.1.3 Aligning short reads back to reference transcriptomes – comparison of available tools

The increase in data available through second and third-generation sequencing requires efficient mapping programs to map the relatively short reads back to references. Mapping of short reads is used in genome re-sequencing, DNA methylation studies, RNA-Seq, ChIP-Seq, studies into structural variants, SNP discovery, differential expression studies and metagenomics [147]. As with *de novo* transcriptome assemblers there is no industry standard; programs are often used as a personal preference since each has advantages and disadvantages. The process can be computationally costly and so there is often a compromise between quality and time when setting parameters to neglect quality scores, limit the number of mismatches, disable gapped alignments or limit the gap length as well as ignoring available SNP data [147].

Several studies have compared mapping programs, especially those most widely used, Bowtie, created by Langmead at the centre for bioinformatics and Computational biology at the University of Maryland and BWA devised by Li and Durbin from the Sanger institute, both of which use the Burrows-Wheeler Transformation indexing methodology [148,149]. These will be explained in greater detail in the following section, after a discussion of the key features of mapping with respect to the nature of sequencing data and current methodologies (Hatem *et al*, 2013).

The chemistry of sequencing reactions, in both first and second-generation technologies, means that the ends of reads are less likely to contain errors and so are used in seeding to improve accuracy [147]. The increased length of reads as second-generation technologies developed, as well as making *de novo* assembly of transcriptomes practical, has also meant that algorithms for seeding and assessing errors in reads have needed development. Mapping tools for Illumina data can use a base quality score, Q, for each base to decide mismatch locations and accept/reject reads based on the sum of these scores at mismatch locations. Q is defined as $-10\log_{10}(e)$, where e is the probability a base being called incorrectly, and was developed from the Phred quality scoring method used with automated Sanger sequencing [147]. Inserting or deleting gaps (INDELS), within limits set by the user, can be used to maximize sequence alignment, but with a penalty of increasing computational time. Computation time is also increased by using paired-end reads, but this increases mapping confidence as the distance between the two ends can be more accurately predicted as well as reducing miss-assemblies [147]. With alternative splicing leading to transcripts that share sequence and the removal of non-coding regions (introns) and joining of the coding regions (exons) the mapping process is further complicated. This is amplified if the sequence covers the exon-exon junction, as the intron length is therefore unknown (can range from 250-65130 nucleotides in model eukaryotic organisms) [150]. Finally as mismatches are often only one nucleotide it is difficult to distinguish if they are genuine mismatches or SNPs and so knowing the location of SNPS helps resolves true mismatches as well as coverage depending on the overall coverage profile of the assembly as discussed earlier. [147].

The first step in any mapping program is the indexing of either the reference or the raw reads using either a hash table or Burrows-Wheeler Transform indexing. The latter is used in the widely adopted alignment programs Bowtie and BWA.

3.1.4 Hash Tables

A hash table is made with the sequence as the key and a list of positions where the sequence is found as the value [147]. Several currently available alignment programs use this method, differing in whether they index the reference or the reads. Four of the most well-known programs that index the reference genome are FANGS (designed to map 454 reads to a reference), NOVOalign (commercial aligner from Novocraft), Mr and MrsFAST (both designed for prediction of copy number variation in duplicated sequences that index k-mers and provide all mapping loci, but without allowing gaps in the case of MrsFAST) and GSNAP (designed by GenenTech Inc) where the genome is split into overlapping 12nt oligomers, sampled every 3 nts so that the location of these substrings can be found and combined [151-155].

Two programs that use a read index are RMAP, (Cold Spring Harbor's Solexa read aligner) and MAQ, (a Sanger Institute and BGI joint project to align shotgun reads to a reference genome) that can be used to scan a genome through multiple hashes of reads [149,156].

3.1.5 Burrows Wheeler Transformation (BWT) aligners

The BWT is a reversible permutation of characters in a text and is used for compression and indexing [157]. It was first created by Burrows and Wheeler in 1994 and is now used alongside the FM index proposed by Ferragina and Manzini in 2000 [158,159]. BWA, Bowtie and SOAP2 (designed by Li *et al* to improve computer memory usage and improve alignment of the SOAP platform) all use BWT [148,149,160]. See figure 3.1. They differ in their approaches to exact and inexact matching. For exact matching (seeds match reference exactly) both Bowtie and BWA use FM indexes of the reference and a modified FM matching algorithm [147]. SOAP2 combines both BWT and hashes to index, to speed up exact matching [147]. BWA has a specific inexact matching algorithm that uses a backtracking method to search for matches between the substrings of the reference and the query sequence within a defined distance [149]. SOAP2 however carries out inexact matching by splitting reads into fragments related to the number of mismatches [160].

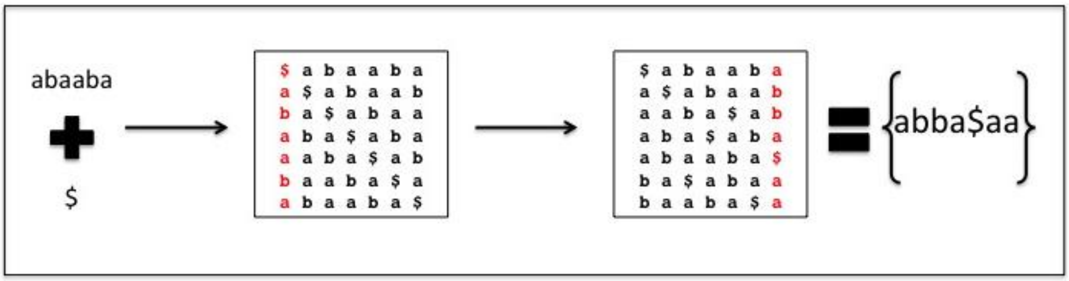


Figure 3-1 Schematic of a BWT.

a) Sequence to be transformed has $ attached to it ($ is a terminator that is not in the sequence and is considered lexicographically prior to all other characters ($<A<C<G<T))
b) All possible permutations of the sequence are stacked vertically starting with $sequence (shown highlighted in red)
c) The rows are then ordered alphabetically
d) The BWT is the final column of the ordered matrix

[148]

3.1.6 Comparison of available aligners

The key considerations when comparing aligners are throughput, memory footprint, and the percentage of mapped reads (both correct and incorrectly mapped) [147]. Aligners such as MAQ and SOAP have a very high computation cost; in order to align the 140 billion bases for the human genome project MAQ would take >5 CPU months and SOAP >3 CPU years [148]. Therefore aligners with this high level of computation cost are not advisable for second-generation projects due to the high number of short reads. Bowtie and BWA are much faster, with Bowtie showing alignment to the human genome at >25 million 35bp reads per hour due to the reduced memory usage required for BWT methods [161]. The memory footprint for the human genome project using Bowtie was only 1.3Gb [148]. Most users keep the default settings on the aligners, which are as follows, along with any user options for both Bowtie and BWA.

Bowtie: the maximum number of mismatches in the seed is set as a default of 2 but can be 0, 1, 2 or 3 with the seed length set to 28. The maximum number of mismatches in the read is based on the read length. These can clearly affect the percentage of mapped reads. The quality threshold is set at 70. Paired-end reads and gapped alignments are allowed in Bowtie analysis and the minimum and maximum insert size for paired-end reads is set to 0 and 250 in Bowtie, or 0 and 500 in the newer version Bowtie2 [147].

BWA: BWA disables seeding. Its maximum number of mismatches per read is also set according to the length of the read. Gapped alignment is allowed, as are paired-end reads, with the minimum and maximum insertion being the same as Bowtie2 [147].

3.1.6.1 Bowtie or BWA

Several studies have been carried out to compare the two methods without either being clearly superior to the other, so the choice of which to use is really a case of personal preference. Bowtie was shown to be more accurate in mismatch experiments carried out by Hatem et al., (2013) while Medina-Medina et al., (2012) recommended both, stating that Bowtie was faster but BWA was more sensitive on the human data tested. In SNP calling, Wong et al., (2012) found that Bowtie mapped more and detected more SNPs than MAQ (BWA's successor) with more false negatives reported for BWA. A study by le Roex et al., (2012) on SNPs in African Buffalo showed similar levels of positive polymorphic determination (Bowtie 43-54% and BWA 57-58% were shown to be real polymorphisms) but fewer reads were mapped overall with BWA. As for data with high level of repeats, Yu et al., (2012) showed that both methods performed equally well with similar false positive rates in repeated regions.

Therefore it was decided that Bowtie would be used in the project since it performs similarly to BWA and can be used alongside VarScan (SNP finding program used in this project).

3.1.7 Analysis of polymorphic differences

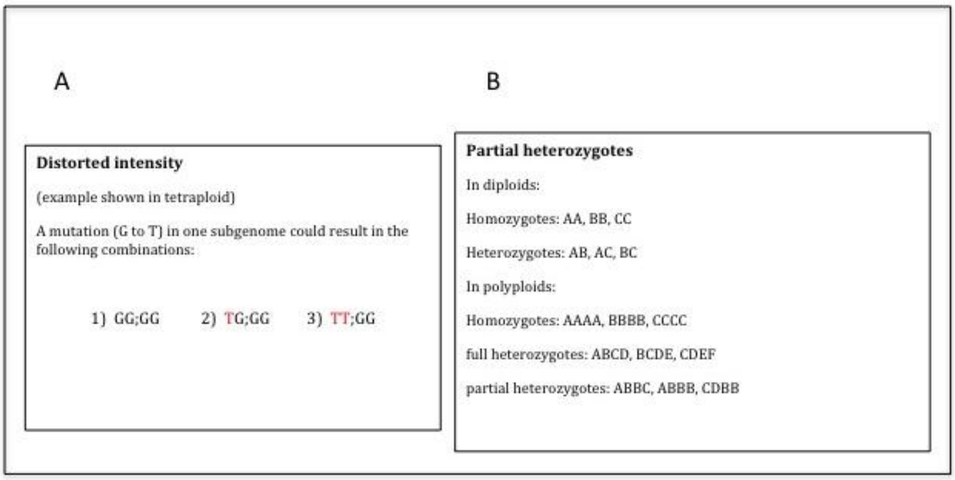
The increase in available data and extensive sequence depth with second and third-generation sequencing technologies has caused a shift away from tradition capillary methods for SNP discovery towards aligners and sequence-based SNP callers [162]. The extensive read depth allows for detection of rare variants and rare transcripts, making it possible to look for transcripts involved in the production of relatively lowly produced molecules such as alkaloids and other secondary metabolites.

The use of transcriptome based molecular markers has the potential to make a great impact on trait linkage studies [111]. The use of transcriptomes can decrease the cost of SNP discovery by allowing for the analysis of the intermediate step between gene and protein linking functionality, by focusing in on the transcribed elements it is possible to look for markers directly linked to phenotypes of interest in areas more likely to undergo recombination [163]. In this project it is hoped that by looking for polymorphic differences between two closely related varieties of *Narcissus pseudonarcissus*, possible SNP markers could be determined linked to galanthamine production. The first step in the process is the analysis of the whole transcriptome before narrowing the search to possible alkaloid biosynthesis related genes (Chapter four).

As with the other techniques discussed in this chapter, the rapid increase in sequence data and methodologies has resulted in a rapid growth and development of novel SNP calling programs, and there is no one method preferred or endorsed by the bioinformatics community [164]. However most current SNP “callers” are developed for use with diploid species and rely on base-calling and mapping quality for sources of error [165]. The methodologies currently employed detect single locus differences in diploids. However due to the genomic nature of polyploids this is not viable [166]. Seventy percent of angiosperms are known to be polyploids so as studies on non-model organisms increase, the need for SNP callers that can deal with varying levels of polyploidy is becoming more apparent [167].

3.1.7.1 Identification of SNPs in polyploids

There are several fundamental biological issues involved in SNP discovery in polyploids. Firstly the difficulty in resolving auto and allopolyploid makes predicting the expected allelic frequencies extremely difficult [167,168]. Ideally, in an allopolyploid if the progenitor diploid species are known for an even numbered polyploid then they can be treated as separate genomes [167]. However the progenitor species are often not known or complicated due to duplication events making SNP/orthologous allelic copies difficult to resolve [167]. The determination of allelic frequency or copy number is also affected by recombination, highly repetitive sequences, in or out breeding and asexual or sexual reproduction [167,168]. Ninety nine percent of apomictic plants are polyploids and often have uneven polyploidy (e.g. triploids). The subgenomic makeup of polyploids also gives rise to distorted frequency and dominance as well as the existence of partial heterozygotes (see figure 3.2) [168].



A

Distorted intensity

(example shown in tetraploid)

A mutation (G to T) in one subgenome could result in the following combinations:

1) GG;GG 2) TG;GG 3) TT;GG

B

Partial heterozygotes

In diploids:

Homozygotes: AA, BB, CC

Heterozygotes: AB, AC, BC

In polyploids:

Homozygotes: AAAA, BBBB, CCCC

full heterozygotes: ABCD, BCDE, CDEF

partial heterozygotes: ABBC, ABBB, CDBB

Figure 3-2 differences between polyploids and diploids that alters copy number prediction.

(*Example shown is tetraploid, not all possible variations or states of heterozygosity are shown)*

A) One mutation within a subgenome can distort the distribution of allelic frequency
    1) A quadriplex T configuration
    2) A simplex T configuration
    3) A duplex T configuration
B) Partial heterozygotes give rise to varying levels of allelic frequencies

[168]

Current SNP detection programs give only one alternative allele; often even those that are capable of dealing with polyploid data still only consider the true SNP frequency as 0.5 in heterozygotes and so miss true SNPs in polyploids [168]. Therefore to gain full understanding of the polymorphic differences between the two varieties, a bespoke perl script pileup_parser.pl was designed to parse the pileup output from the SAMtools mpileup tool that is used by some SNP callers. This could be used alongside VarScan, another SNP caller, to look at the exact differences between the varieties. VarScan was chosen because it could take the same input file as pileup_parser.pl and was compatible with several aligners such as Newbler, Bowtie and Novalign unlike most SNP calling programs that only work with one specific aligner [162]. More over the most recent version of VarScan (version 2) allows for SAM/BAM inputs making it a more universal caller [169]. The VarScan output can easily be compared with the output from the bespoke perl script and has been shown to perform well on both 454 and Illumina data, allowing for easy determination of polymorphic differences from both the 454 and Illumina datasets acquired for *Narcissus pseudonarcissus* [162].

3.1.7.2 VarScan

VarScan can be used on both individual and pooled samples and has been tried and tested on both 454 and Illumina data. It is widely used as it takes the direct output of many read aligning programs [170]. The pileup format output from SAMtools is analysed and the alignments scored and sorted on a per read basis [162]. It is able to report SNPs and INDELs with their corresponding chromosome/contig co-ordinates, alleles, flanking sequence and read counts [162]. The best alignment for each read is screened for sequence differences after the removal of low identity and/or ambiguous alignments. It takes into consideration coverage, quality, variant frequency and the number of reads that support each alignment and the user can adjust these parameters. In a study of 454 data comparing VarScan and Newbler, VarScan was able to correctly predict over twice the number of SNPs (59.78% compared with 22.28%). It showed comparable results in a comparison with MAQ using Illumina data (97.21% compared to 94.71%, although only 3 of the SNPS passed the MAQ SNPfilter)[162]. This study shows a clear difference in using 454 or Illumina reads, a large number of the SNPS (60% compared to 97%) were not found in the 454 data, this might have been due to the pooled nature of the Illumina data as well as the improved coverage (70X compared to 125X). The new version of VarScan, version 2, has several key improvements including:

1. Input can now be in the SAM or BAM file format allowing for compatibility with the pileup formatted output from the mpileup program of the SAMtools suite that uses Bowtie as an aligner.
2. Increased performance and ease of use since it now runs in Java and so can be used with any operating system.
3. In addition to detecting variants, VarScan 2 can call consensus genotypes based on real counts and allele frequencies.
4. Detects exome-based copy number alteration.

169

3.1.7.3 Pileup_parser.pl perl script for determining all possible allele variation

Both VarScan and the pileup_parser.pl script quantify the variation between the reference and the short Illumina reads by using the pileup file from SAMtools mpileup. SAMtools mpileup uses the output from Bowtie in a sorted BAM file format and the reference as a fasta file; an example pileup output format can be seen in figure 3.3.

```
comp32869_c0_seq1       7       G       624
.................................................................................
.................................................................................
...........................+1A...................................................
.................................................................................
......................,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,.......
.................................................................................
....................,,,,,,,,,,,................................................,,
,,,,,......,,................,,,...................,,,^K.^K.^K.^K.^K.^K,
FBFFFFFFDFFFFFFFFFFFDFFDFDA@FFFFFF0FFFDFDFFFFFFFFFFFFFFFDDFFJFFDFFFFFFFFFFFFFDDFDFFFFFD
FFDFFFFFFFDFDFFFFFFFDFFDFFFFF;FFFFFFFFFFFFDDFDFFADFFFDFFFFFFFFJDFFFFFFF?DFFF=FFDFDFBF
BBFFFF=FFFFFFFFFF<?FFFFFFFFFFFFDAFDFDFFFFFFFFFFFFFDDDFFFFDDFFDFFFFDFFFFFFFFFFDFFFFFFFF
FFFFJFFFD;FFFFFDFFFFFFFFFFFFF6AFFFFFDDFFFF@FFFAFFFFFBFFFFFFFFFFDFFDFFDFFF:FDFFFFDHD
FFFFFFFFFFFFFDFFFFDFDDDBDDIADADDBDDDBAB<DBHJDDDBDBDDDDDDADDDID?D<DBDJBJJF0.@/:13:<
;B<0>=;5/@;8=10;<18=?/FE2=>>>;=>./9F?.:6C<?<;;<?B...<;></<?8<;@<;;?<;;<9F<1;4=<0>
/F;;<.<::0;8<;<=></1:D/.<;/;:A==1<1<9111../=14<=@<.=515111@.=>1;3/.111=../1;:BD11
@31>10>1>9>>7>>>>>1>>1>>1>>1>>>>>4<4>>>>>>>>>>>>>>>3CCCCCB
comp32869_c0_seq1       8       T       647
.................................................................................
.........................................................-
1G............................................A..................................
.................................................................................
...............................,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,...............................................................................
.................................................,,,,,,,,,,,,,,,,,,,,............
................C........................,,,,,,,.........,,................,,,...
...............,,,.....,^I.
:2<@HAF0D<DBAD<<DC0DAAD==ABBA<<FA<D=C:DD@CC=<<ADDAD=AD<J@D?AFD2=@AHADA@D00C<DDD<<
BEA<@A=D<DD@F@@ADAD?F?0DBD<F<AD@=FAHA=A<DD<F?AD@DAFAF:DBD@EEA@0D<A2AF6D0A3DB2DF=D
0AHC==<A2F<00A<A<2A<HD=D<2DD@BEA0D@<FA<D<<D<@B<ADF0<A@?ABADF?A0<DD@C@<=@8FA2@FG<2
=DBDB@AADB8AAAFD<ABB220DH0DD=AA=8@ADDAA@AA?DFD@DDFAAD@D=0D<D<@D=A8D@FAAAB==<AD@8A
@DDDBDDCDDDG9DGDDDDCD??DCD?HJDDDDC?DDDDDD:DDDGDDD9D<DJDJJ@=BD@2D=AAA=A@<@A><=@A<D
AFDF?3<AA<F@AB=A2A@A==AB<D<B0DB<D<D5.B<4F@B@A012<621A?=A222A=A<A@0=@AA@A@<D?A>0DD
@@CA>@AA=1<A<D0<=?7<032CDD>D5DDDCDD56CAD50AA<AD>F?:@.;>A=10=BH2BBA@1<5C/@>>A.1ABD
AD3/1=>/;<112=AACDCDDDFB==20B=AGDBA64ABABADBDABABC;DB8?=@@BB??==?B:=BBCD2CCC@@D@
comp32869_c0_seq1       9       G       672
.................................................................................
.............................................................*...................
.................................................................................
.................................................................................
..................,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,.............
.................................................................................
........................................,,,,,,,,,,,,,,,,,,,......................
................................,,,,,,,,,..........,,..................,,,.......
............,,t.....,.^).^K.
A@CFHDH@H?HFDHCCFF@HFAHA@FFFFCCH7DCBCFCHHFFFCCCFFFDHC?H@JCH?FHHCCDDHFHDDH<CCFCFH<
CC0FHDCFCCFCHDDHDDCFD@F?HD@DFHCHCFDDCHF<BDAFDHDCFFDHDCDHFHCHFHDHHF@@FCDADH:H@D7FF
<HHCG0FHFCCCF<H@@0DCFCAFCHHDHCAFHDFHC@8?CHFAFCDFCFFDFFFCCFCFFFFHHFF?CHHD7FFDCD@HD
CDFGCCCD6FFFDDDHDCDFDHFCFFFA<@HH@HD<CDDCCFDH6HFDDF0FFFH@FDDHHDFHFHD@HD0HCDHCDC@AF
HFFDFDDCFHF<FDHHBFDDADDDGDCDDDBDDB9D9D8HIDDBBD8DDDDDDBDDFDDD=D8BJDGJDAFHD5FBFDDCD
CBFFA@BCDCFDHDDB6CFDCHBCDCF<DDDDADDCDDD1FFCGACF80F>5FDADDG@D=A74A<D@AD5A7<ADCF@DD
AADFHCFD=HDDH@DFD@HDADDFD>CFCD3CBD2D<@@C?DDAD7DDDBDB78DDD39DDFH?HADD1=AB>>3>.DHAD
FDD2D=H5AA?D/2HDDDHA22A@3=DA>5AFDADB.DBDDFDAB55DADDGDD6D5=DD<DDDDDDBDFDB?CDDDDF=F
DDDDDDDDBDDD@D0CCCCCDC1C
```

Figure 3-3 Example of SAMtools pileup output format.

The tab-separated values represent the contig, base co-ordinate, reference base, number of reads, read bases and base qualities.
. match to reference on forward strand
, match on reverse strand
A/C/G/T/N/a/c/g/t/n show alternative nucleotide “mismatch”
+ followed by a number and letters indicates an insertion
- followed by a number and letters indicates a deletion
^ start of read segment
$ end of a read segment

For simplicity, as no open reading frame information was available and only SNP information was required, the script was written to discount $, ^ as well as

INDELS as these should be relatively rare in transcriptomes due to the frame shifts caused by INDELs that could prevent the transcription of the genes. The script parses the pileup output to give information on the contig ID, position, reference, number of reads that support an A,T,C,G,N,a,c,t,g or n nucleotide, the total number of reads, the percentage of reads that hit A,T,C,G,N,a,c,t,g or n, the total percentage and total percentage of SNP if present. An example output is shown in figure 3.4. The implementation of the script and its results are discussed in sections 3.3.2, 3.4.2.6 and 3.4.2.7. The script pileup_parser.pl can be found in the folder "scripts" on the Appendix disc.

```
comp32869_c0_seq1       8       T       1       645     1       1       2
650     0.153846153846154       99.2307692307692        0.153846153846154
0.153846153846154       0.307692307692308       100     0.461538461538462
comp32869_c0_seq1       9       G       0       1       0       670     1
672     0       0.148809523809524       0       99.7023809523809
0.148809523809524       100     0.148809523809524
comp32869_c0_seq1       10      T       4       670     3       5       6
688     0.581395348837209       97.3837209302326        0.436046511627907
0.726744186046512       0.872093023255814       100     1.74418604651163
```

Figure 3-4 Example of output from pileup_parser.pl script.

Parses the pileup file to give contig ID, position, reference, number of reads to support A, T , C, G or N, total number of reads, percentage of reads for A, T, C, G or N, total percentage and percentage of SNP if present.

3.1.7.4 *The use of pileup_parser.pl to predict ploidy levels at the nucleotide level*

The perl script is also used to predict ploidy level within the data. Sequential examination of the percentage of every nucleotide at every position within a sequence can be used to identify alternative alleles and their percentages of reads will determine the ploidy level at each locus. For example, in an "ideal" tetraploid, at any locus each alternative allele would have 25% of the total reads, while there would be 33.3% for triploids. The perl script takes percentages at each position and predicts "triploid" if three nucleotides are represented at over 20% with the final nucleotide representing less than 10%. Tetraploid is predicted if all four nucleotides represent over 20%. Those loci that have no alternative allele are labeled "same" as well as those with nucleotides representing less than 15% each. This is an estimation as it does not take into account any quality data or total number of reads and so is only used as a guide.

3.1.8 Analysis of transcript level differences between individuals from two varieties of *Narcissus pseudonarcissus*

A further tool available for investigating secondary metabolite production in plants is the analysis of transcript expression. Differential expression (DE) is used in a wide variety of studies including identifying differences between tissues, studying developmental changes and microRNA target prediction. This is discussed in detail in the introduction chapter section 1.5.7. The first step is a transcriptome wide analysis of differential expression. In order to look at differential expression it is important to acquire accurate estimates of expression while taking into account both technical and biological variation [171]. However due to both financial constraints in this study, as well as sample availability, it was not possible to carry out repeats on the Illumina sequencing and so in this project the difference between two individuals from the two varieties will be compared and used to predict variation that can be analysed further using qPCR in numerous samples from different individuals of each variety.

It was decided that since the project aimed to discover putative genes involved in galanthamine production, and variation within this subset of data, an initial bioinformatics analysis would be used. The method must be able to accommodate two different varieties (conditions) and RNA-Seq data and therefore Bayesian inference of transcripts from sequencing data (BitSeq) was chosen. This approach is also compatible with Bowtie, which was the aligner of choice in this project.

3.1.8.1 BitSeq

In theory, with correct sample preparation the number of reads aligned to a given gene is proportionate to the abundance of fragments or transcripts for that gene. However splicing occurs during transcription resulting in multiple transcript sequences that share the same genic sequence [171,172]. Since the origin of these shared sequences can be difficult to determine, expression must be estimated in a probabilistic way. BitSeq uses a Bayesian approach [171], improving on previous methods by combining a probabilistic model of read generation with a MCMC (Markov Chain Monte Carlo) algorithm for Bayesian inference over the model. This allows for the consideration of uncertainty unlike previous methods such as the expectation-maximization approach that only resulted in a point estimate of transcript abundance. Other methods using MCMC and Bayesian approaches do not allow for the multi-alignment of transcripts permitted within BitSeq, which thus gives a better representation of the transcriptome [171]. Overall, BitSeq incorporates read and alignment quality, adjusts for non-uniform distribution and can handle paired-end reads for DE between two conditions, or plant varieties[171].

3.1.9 Transposon element variation within plants

Initial work on the assembly of the raw Illumina data in this chapter showed unusually long contigs for *de novo* assembly (>5000bp, with maximum length 30656bp and a range of coverage from 1.93 to 7432.51X) suggesting further filtering was required. A personal communication with the Trinity group suggested that the longer sequences might be due to transposon contamination, or concatenated sequences.

As discussed in section 1.5.2 of chapter one, plants show great diversity in genome size ranging from 64 Mbp (*Genlisea margaratae*) to 150 Gbp (*Paris japonica*) [173]. The difference is not only caused by polyploidy events but also the amplification of transposons and related sequences [173]. Transposable elements (TEs) make up 15-84% of plant genomes and influence evolution via increasing genome size and altering gene function [174]. Examples of this influence can be seen throughout the plant world; in maize genome size changes are linked to long terminal repeat (LTR) TEs, the sunflower (*Helianthus annuus*) genome contains 81% TEs (77% LTR) and the genome of wild rice (*Zizania palustris*) has doubled without a ploidy change due to expansion of *Copia* and *Gypsy* TEs. TEs are also linked to the varied genome size seen in species of grass [173,175].

There are two main classes of TE in plants, those that transpose through an RNA intermediate such as LTR/*Copia* and LTR/*Gypsy* and those that move through a

DNA intermediate [174]. TEs vary greatly in copy number and sequence between closely related species and within the same species as well as varying in heterogeneity. The reverse transcriptase domain has been shown to be between 5-75% heterogenic [174]. These repeated sequences could cause issues with *de novo* assembly since the sequence can cause branches in de Bruijn graphs and lead to the assembly of chimeric sequence. Therefore, as the number and the level of activity of TEs in daffodils is unknown, and the genes of interest are likely to be rarer transcripts, the removal of such repeated TE reads would be beneficial [173].

There are several methods available that could be suited for this purpose;

3.1.9.1 TransposonPSI

TransposonPSI is a PSI-BLAST application from the Broad Institute used to identify homologues of protein or nucleic acid sequences from diverse families of TEs [176]. It contains a database of TEs that include gypsy and Copia polyproteins, cryptons, helitrons, numerous DNA transposon families and LINE retrotransposon orfs [176]. It produces two output files, allHits that contains all possible PSIBLASTP matches and topHIts that gives the best hit. As this covers a wide diversity of TEs it was used since no TE information is currently available for *Narcissus*.

3.1.9.2 TREP

Another database with potential for identifying TEs is the TREP database of the International Triticeae Mapping Initiative (ITMI) [177]. This is an extensive database that contains repetitive DNA sequences from different *Triticeae* species. As this is well annotated and *Triticeae* well studied it is hoped that this database will be extensive and could lead to TE discovery in *Narcissus.* This database could simply be used to create a BLAST database and the *Narcissus* sequences compared for homologues, although to run BLAST on raw reads is inefficient.

3.1.9.3 BMTagger

An alternative method would be to use the program BMTagger (Best Match Tagger), developed as part of the NCBI Human microbiome project to remove human sequence contamination [178]. As the program requires an indexed reference of the human genome to match against it is possible to submit any indexed reference such as the TREP database. BMTagger has two key steps. First, bmfilter classifies reads as contaminants by looking for 18mers that match the reference. This step does not require alignment and so it much faster than BLAST or BWA [178]. If sequences are not classified during this step an alignment to the reference with up to two errors allowed is attempted, Any alignments found in the 1st mate (of paired end reads) also results in the removal of the 2nd mate [178].

As there is no precedent for this sort of work in *Narcissus* TEs will be identified using TransposonPSI, BLAST against TREP and BMTagger against TREP and the results compared to find the most stringent method. (See sections 3.3.5 and 3.4.2.10)

3.2 Aims and Objectives

To produce and analyse high depth short-read sequence data of two daffodil varieties with known differences in alkaloid levels, RNA from the basal plate of one individual from each variety was sequenced on the Illumina Hi-Seq platform for comparison to the 454 references created in Chapter two.

The reads were assembled *de novo* to account for any transcripts missed by the lower coverage 454 reference as well as mapped back to the reference and back to the assembled Illumina data to evaluate transcriptome wide variation between the varieties. Differences were determined in both SNP calling and DE analysis.

3.3 Methodology

3.3.1 Illumina library preparation, sequencing, assembly and annotation

3.3.1.1 Extraction of RNA from N. pseudonarcissus var Andrew's Choice

The CTAB method (Chang *et al.,* 1993) followed by Qiagen RNeasy Plant kit RNA clean up was tested for the Andrew's Choice variety. Unfortunately this method showed varying results and so a trial was conducted of alternative methods of RNA extraction. Several RNA extraction kits from commercial sources were compared with each other and the CTAB method (see Table 3.1). All methods were tested using frozen tissue of Andrew's Choice from 2011 field trial replicates that were not otherwise needed for downstream processing.

Table 3-1 Methods used in RNA extraction trial for Andrew's choice.

All kits were used as per manufacturer's instructions unless stated in the methodology column.

Methodology	Manufacturer	Catalogue Number
CTAB followed by Qiagen RNeasy clean up	CTAB Chang et al., 1993, Qiagen	74903
MoBio Power Plant RNA extraction kit followed by RNeasy clean up with optional DNase on column step. Optional step during MoBio protocol of adding 50µl of PSS (phenol separation solution) was included	MoBio and Qiagen	MoBio: 13500-50, Qiagen: 74903 and 79254
MoBio Power Plant RNA extraction kit with optional PSS usage.	MoBio	13500-50
MoBio Power Plant RNA extraction kit with optional PSS usage. Followed by Mo bio RTS DNase.	MoBio	13500-50 and 15200-50
MoBio Power Plant RNA extraction kit with optional PSS usage. Followed by Qiagen RNeasy clean up.	MoBio and Qiagen	MoBio: 13500-50, Qiagen: 74903
InnuSPEED Plant RNA kit using Mo biolyzer for homogenisation (two 45 second cycles at 4200rpm) [using MoBio Powerlyzer™ 24 bench top bead-bead homogenizer cat no- 13155] and PL lysis solution.	Analytik-Jena	Supplied by Web Scientific to test. No current catalogue number
InnuPREP Plant RNA kit using PL lysis solution. Lysis (20 min, room temperature) vortexed every 2-3 min.	Analytik-Jena	Supplied by Web Scientific to test. No current catalogue number

3.3.1.2 Illumina cDNA library preparation for N. pseudonarcissus var. Carlton and N. pseudonarcissus var. Andrew's Choice

RNA from two different Basal plate samples of Andrew's Choice and two different basal plate samples of Carlton were utilized to produce RNA-Seq libraries. The initial RNA samples were analyzed using an Agilent RNA 6000 Nano Chip (see section 3.4.2.1). The rRNA from the four samples was then removed using the Epibio Ribo-Zero™ rRNA removal kit (plant seed and root) following the manufacturer's instructions with step 3.d, being carried out using the Agencourt RNACleanup up kit.

The four samples were then used to produce four ScriptSeq™ RNA-Seq libraries using the Epibio® ScriptSeq™ v2 RNA-Seq Library Preparation Kit following the manufacturer's protocol. In order to sequence all samples in one pool, ScriptSeq

Index PCR Primers were used at step 4.e. (see table 3.2). Step 4.f of the protocol was carried out using the Agencourt AMPure XP kit. The four resulting libraries were analyzed using both a Quibit fluorometer with RNA assay kit and an Agilent Bioanalyser Nano High Sensitivity DNA chip (see section 3.4.2.2).

Table 3-2 ScriptSeq Index PCR primers used in library preparation.

Sample	ScriptSeq Index PCR Primer	Primer sequence
Andrew's Choice 1	Index 3	GCCTAA
Andrew's Choice 2	Index 4	TGGTCA
Carlton 1	Index 5	CACTGT
Carlton 2	Index 6	ATTGGC

3.3.1.3 Analysis of Sequence data

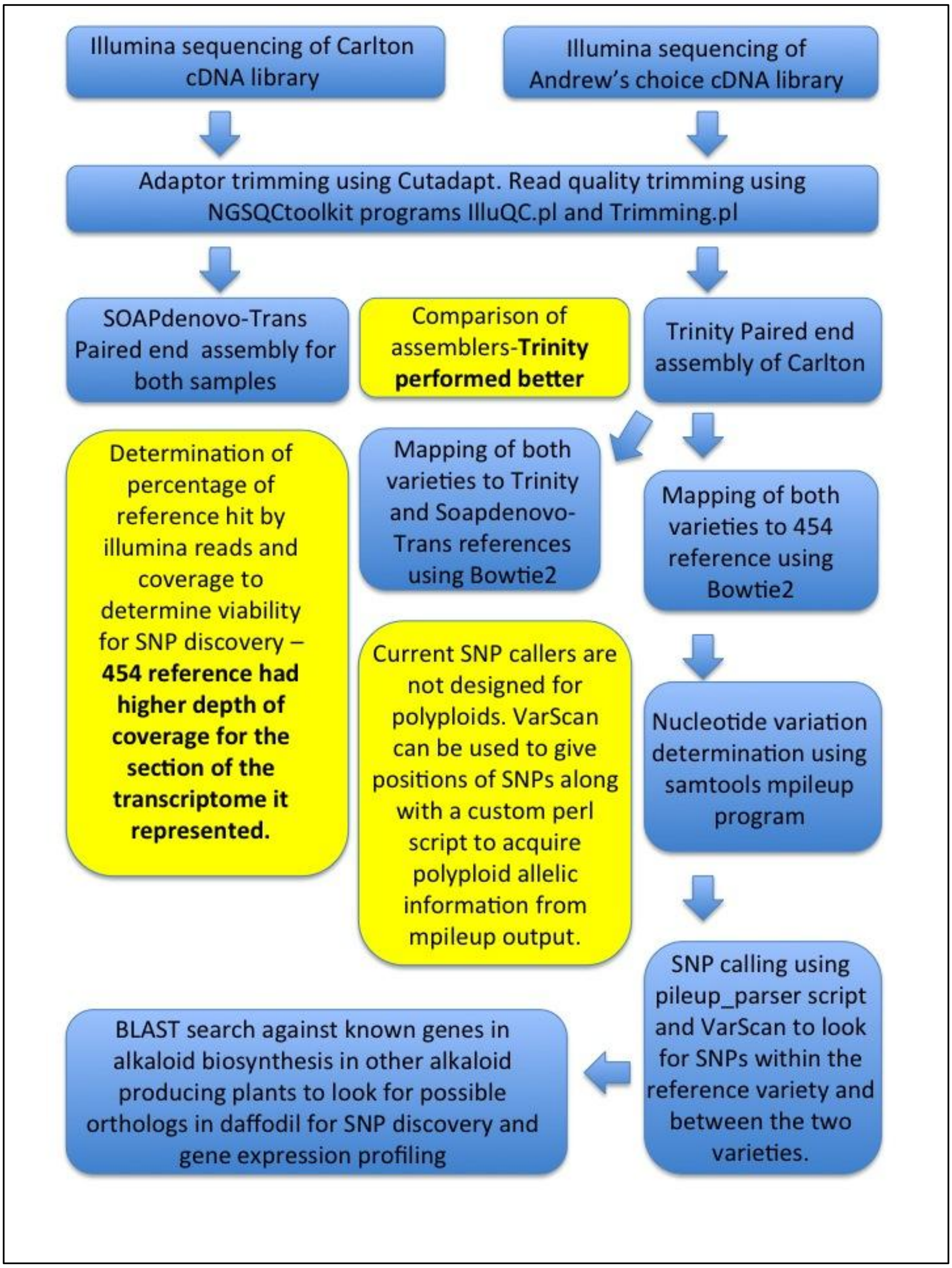

Illumina sequencing of Carlton cDNA library
Illumina sequencing of Andrew's choice cDNA library
Adaptor trimming using Cutadapt. Read quality trimming using NGSQCtoolkit programs IlluQC.pl and Trimming.pl
SOAPdenovo-Trans Paired end assembly for both samples
Comparison of assemblers-Trinity performed better
Trinity Paired end assembly of Carlton
Determination of percentage of reference hit by illumina reads and coverage to determine viability for SNP discovery – 454 reference had higher depth of coverage for the section of the transcriptome it represented.
Mapping of both varieties to Trinity and Soapdenovo-Trans references using Bowtie2
Mapping of both varieties to 454 reference using Bowtie2
Current SNP callers are not designed for polyploids. VarScan can be used to give positions of SNPs along with a custom perl script to acquire polyploid allelic information from mpileup output.
Nucleotide variation determination using samtools mpileup program
SNP calling using pileup_parser script and VarScan to look for SNPs within the reference variety and between the two varieties.
BLAST search against known genes in alkaloid biosynthesis in other alkaloid producing plants to look for possible orthologs in daffodil for SNP discovery and gene expression profiling

Figure 3-5 Schematic of Illumina data analysis.

The above steps were implemented using a variety of RNA-Seq analysis tools and custom scripts. The steps are described in detail in the following sections.

3.3.1.4 Basic script and program running

All perl scripts and programs were run via the command line on an 8 core (dual 4-core 2.13GHz E5506) Intel Xeon machine with 48 GB RAM and 1TB scratch (2x 1TB disks, RAID1)) unless otherwise stated. The command line prompts can be seen in appendix section 6.10.

3.3.1.5 Trimming and Filtering of raw reads

The raw Illumina reads were trimmed to remove the ScriptSeq Index PCR Primers used for library preparation. The Centre for Genomic Research at the University of Liverpool removed the primers using Cutadapt, an open source program that removes adapter sequences from high-throughput sequencing data [179]. Cutadapt was implemented using the default settings.

The reads were then filtered and trimmed depending on quality scores. The NGS QC Toolbox version 2.3 (http://59.163.192.90:8080/ngsqctoolkit/) was used to both filter and trim the reads. The first filtering step was carried out using the illuQC.pl script in NGSQCToolkit_v2.3/QC that filters reads on the percentage of bases that have a given quality score. The script was run with the paired end option so that both reads were removed if one fell below the given parameters. The filter for adapters and primers was turned off as the adapters were already removed. The percentage of bases that were required to be above the set quality score (20, default setting) was also used at the default setting of 70%.

The filtered reads were then trimmed using Trimming.pl in NGSQCToolkit_v2.3/Trimming. This script trims low quality bases from the 3' end of reads and removes reads below a set read length. The script was run as with the default settings except the minimum quality score was increased from 20 to 30 in order to have a more stringent removal of low quality reads.

3.3.1.6 Comparison of de novo assembly using Trinity and SOAPdenovo-Trans

Two methods of *de novo* assembly were used the first method was Trinity and it was run as a paired end assembly with the system memory usage for the jellyfish step set at 54G as this was the highest accepted level for the system used. The number of CPUs was set as the suggested default of 16.

The second method was SOAPdenovo-Trans-127kmer version run as a paired end assembly using the default settings for both read and scaffold assembly with the average insert size set to 215 as this was the average for the raw reads.

3.3.1.7 Mapping of trimmed Illumina reads to the Trinity assembly, SOAPdenovo-Trans and the 454 references

Bowtie2 was used to align the raw Illumina reads of both Carlton and Andrew's Choice to the three assemblies. Bowtie2 was run on all assemblies using the following command line prompts. The default settings of an end-to-end alignment with the default minimum score threshold of -0.6 + 0.6 * L, where L is the read length. The default alignment mode is also used that searches for multiple alignments and reports the best one.

A comparison of the resulting alignments was carried out by analyzing the percentage of reads aligned compared to the percentage of the reference hit by the reads, and the coverage of these alignments. In order to do this a script, coverageStatsSplitByChr_v2.pl (written by Kevin Ashelford and modified by Laura Gardiner at the University of Liverpool CGI), was used to work out the coverage of each read and the percentage of the reference hit. This script requires a sorted BAM file as input and so SAMtools was used to convert the SAM files from Bowtie to sorted BAM files. An average of these were then determined to compare the three assemblies.

The coverageStatsSplitByChr_v2.pl script was run and the results analysed to give the percentage of reference hit and average depth of coverage using awk commands. See appendix section 6.10 .

3.3.2 SNP calling using custom SNP caller for polyploid samples and the prediction of ploidy level

As discussed in section 3.1.7.1 and 3.1.7.2, the version of VarScan used only considers SNPs with 0.5% heterozygosity and so does not give full details on polyploids. It also does not take into account a second or third alternative allele when determining the percentage of total reads that represent each allele, for example if at a single loci the reference is C and 30 reads are C, 30 are A and 30 are T the percentage is only worked out for each allele as a percentage of 60 not 90. Therefore the pileup_parser.pl script was used along side the VarScan results to pull out the true percentage of reads for each possible allele. This information was then used to analysis possible non-synonymous SNPs as is shown in chapter four. VarScan was run with a minimum read depth of 20, a minimum of 100 supporting reads to call a variant and an average read quality cut off of 20

The perl script was also used to look at polyploidal variation at the individual loci level and so after it was run the results were parsed to count the number of "triploid" and "tetraploid" labels for each loci.

In order to determine if the SNPs found were seen in both varieties, between the varieties or only in Carlton the results were compared on position of SNP, reference and alternative allele suggested. The main analysis to look for non-synonymous SNPS in putative genes (chapter four) was only carried out on SNPs thought to be between varieties.

3.3.3 Transcript level differences determined using BitSeq

BitSeq was used to determine differential expression of transcripts between the two varieties of daffodil. The program requires a SAM file, which is generated via mapping to the reference transcriptome using Bowtie. An index was first creating using the reference transcripts from the Carlton data. The default setting for the step were altered from 5 to 2 for the number of Burrows-Wheeler rows marked and from 10 to 12 for the number of characters to use to create the look up table for the initial Burrows-Wheeler range determination. These were changed as the BitSeq example analysis suggests that these settings work better for data of this nature.

After this step, Bowtie is used to map the reads to the reference index with a maximum number of mismatches of 3 without trimming the low-quality end of each read pre alignment as this was already carried out during filtering and trimming and would increase computational time. At this stage BitSeq can then be run following the example from the BitSeq wiki, http://code.google.com/p/bitseq/wiki/BitSeq.

The resulting files from the BitSeq analysis consist of a file with the PPLR data and a separate file with the contig IDs. Using a simple combination of command line prompts, the first 8 lines (headers and description of data) were removed and the data matched up with the contig IDs. The file was then altered to make any tabs spaces for ease of use in later steps. The perl script tabtospace.pl can be found in the appendix section. The resulting .pplr file could be used alongside a simple split script that pulls out those contigs with significant differences between the two conditions (varieties), that is to say those with a PPLR of >0.95 and <0.05. The testsplit_bitseq.pl script can be found in the scripts folder on the appendix disc.

3.3.4 Annotation of Trinity assembly

The assembled Illumina data was annotated using the same pipeline used for the 454 data (section 2.3.4.5).

3.3.5 Removal of transposon elements from raw reads – comparison of three methods

Three methods were tested to determine which removed the greatest number of raw reads.

3.3.5.1 BMTagger

This program is described in the introduction of this chapter. The program was run using the default settings. The resulting files were then used with a perl script (Dr J Kelly, University of Liverpool) to extract the reads that matched TREP from the original reads ready to re-assemble the reads without transposon contamination. The script is in the appendix section.

The resulting assembly can be compared to the assemblies from the other methods.

3.3.5.2 TransposonPSI

TransposonPSI.pl is a perl script and so was run as written. Only the top hit file was used in the analysis. The output file was filtered and the used alongside the script by Dr Kelly to remove the transposon elements from the filtered and trimmed reads ready for assembly.

3.3.5.3 BLAST against TREP

The format converter FastqtoFasta.pl from NGStoolkit was used to convert the raw files to fasta files then the BLAST (version 2.2.27+, scoring matrix Blosum62) search was run against the 2008 TREP database with the output set to tabulated for ease of comparison and annotation with the search limited to one hit per transcript. The resulting BLAST output was parsed using a bespoke perl script for an e-value cut off of e^{-}). The perl script remove_low_scoring_blast.pl can be found in the scripts folder on appendix disc.

3.3.6 Re-assembly of reads post TE removal

BMTagger results were used for re-assembly using Trinity version 2012-10-05 using the same settings as the original assembly for comparison.

3.3.7 Annotation of the re-assembled data

The new Trinity assembly was annotated using the same pipeline as described in section 2.3.4.5.

3.3.8 SNP and transcript level differences re-evaluated post transposon removal

The SNP and transcript level differences were re-analysed using the new assembly as described in sections 3.3.2 and 3.3.3 and compared to the original analysis.

3.4 Results

3.4.1 Extraction of RNA

3.4.1.1 Andrew's Choice

The results of the trial can be seen in the Appendix, table x. It was clear that the best method was the Analytik-Jena InnuPREP plant RNA Kit with the PL lysis solution. The kit gave consistent high yields and consistent 260/280 ratios. Therefore this method was used alongside a DNase step (Qiagen DNase kit) for the Andrew's Choice variety. The two samples from two different bulbs grown in the same conditions had total concentrations of 696 ng μl^{-1}and 474 ng μl^{-1} pre DNase clean up. Both samples had 260/280 ratios within the range (1.8-2.3) suggesting pure RNA and 260/230 ratios indicating no protein contamination. The samples were stored at -80°C until needed for cDNA library preparation.

3.4.1.2 Carlton

The same method as for Andrew's Choice was used to prepare two samples of Carlton RNA from two different bulbs grown under the same conditions as the Andrew's Choice. The resulting concentrations pre DNase clean up were 997 ng μl^{-1} and 687 ng μl^{-1}. These were also stored at -80 °C until needed for cDNA library preparation.

3.4.2 RNA-Seq library preparation

3.4.2.1 Total RNA analysis

The four samples were analyzed for total RNA concentration using the Agilent RNA 6000 Nano Chip. The results can be seen in table 3.3.

Table 3-3 Total RNA Nano Chip results for the four total RNA samples for consideration for library preparation.

Sample	Concentration (ng μl^{-1})	28S/18S ratio	RIN
Andrew's Choice 1	416.69	1.80	8.1
Andrews' Choice 2	465.16	1.99	8.6
Carlton 1	398.58	1.61	8.6
Carlton 2	466.81	1.74	8.6

From table 3.3 it is can be seen that all four samples have RIN > 6, which is the degradation threshold, suggesting good quality RNA. The 28S/18S ratio confirms this because they are all close to the optimum of 2.0 obtained in non-degraded RNA. This is further demonstrated in figures 3.6-3.9 by the clear single peaks at 18s and 28s and the relatively smooth graphs. All four samples were carried forward to library preparation.

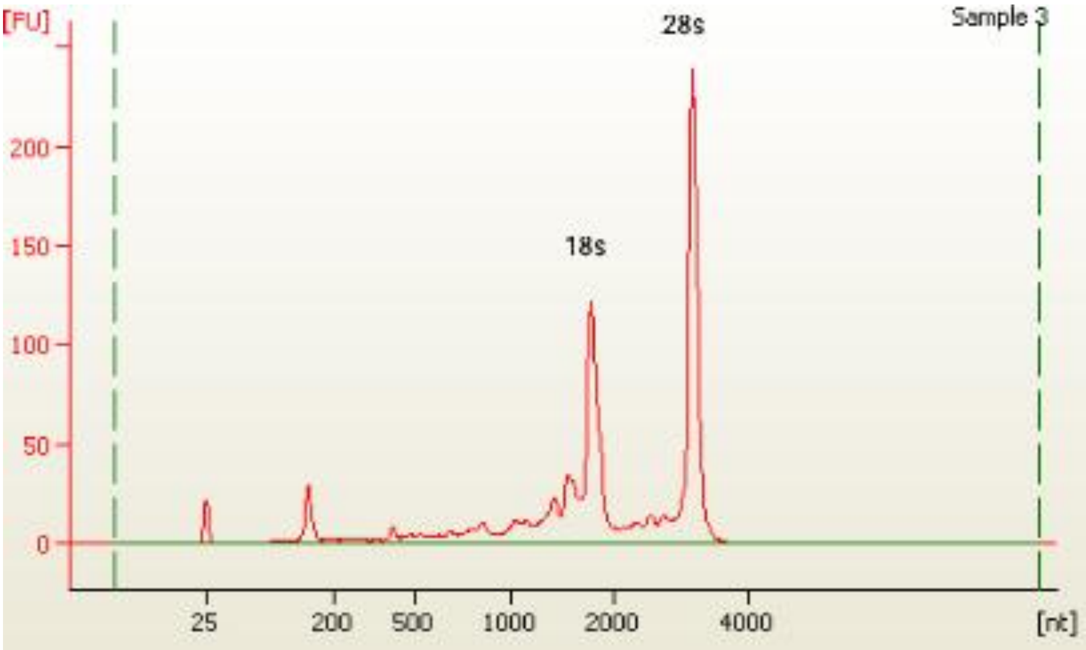


Figure 3-6 Total RNA electropherogram from Andrew's Choice sample 1

The rRNA ratio (28S/18S) was 1.8 with a RIN of 8.1 and a final concentration of 416.69 ng µl^{-1}.

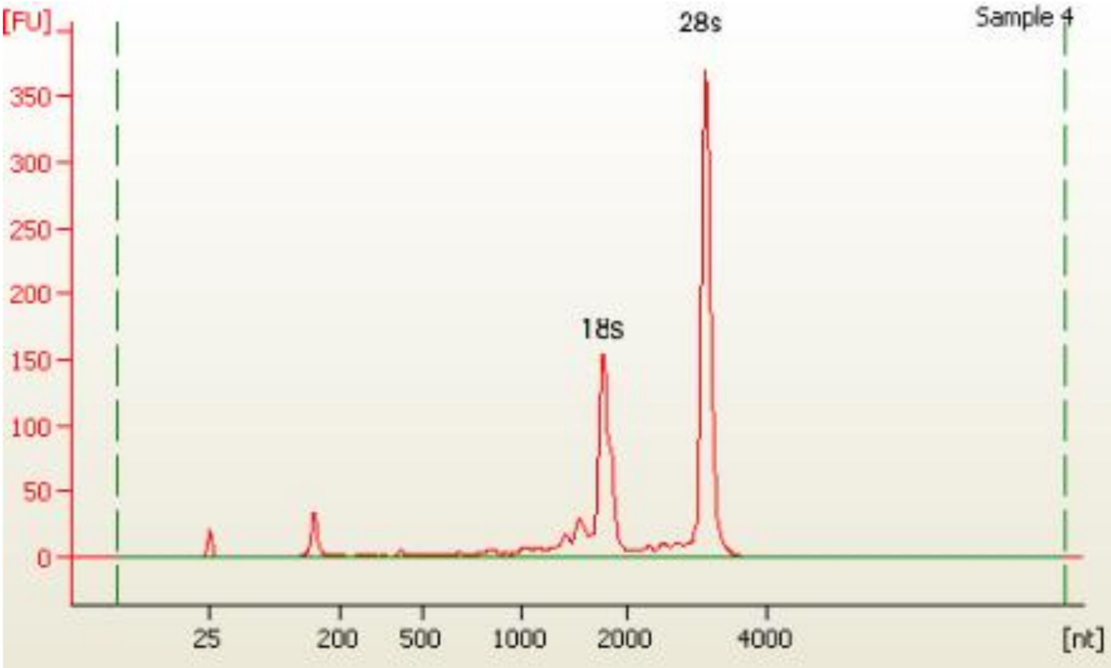


Figure 3-7 Total RNA electropherogram from Andrew's Choice sample 2

The rRNA ratio (28S/18S) was 1.99 with a RIN of 8.6 and a final concentration of 465.16 ng µl^{-1}.

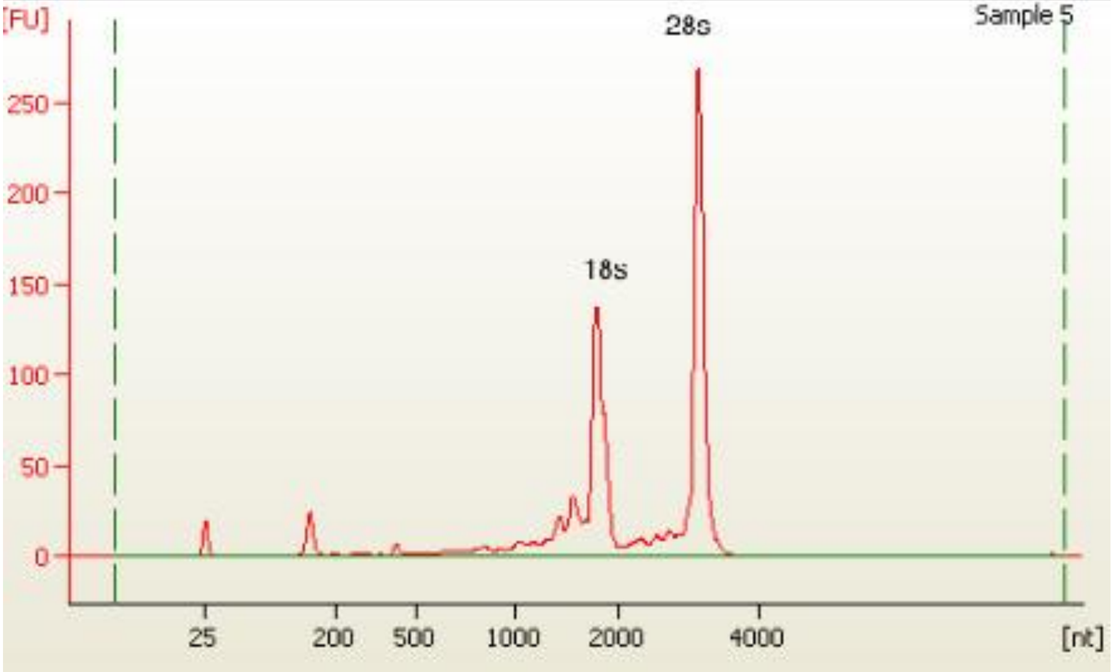


Figure 3-8 Total RNA electropherogram from Carlton sample 1.

The rRNA ratio (28S/18S) was 1.61 with a RIN of 8.6 and a final concentration of 398.58 ng µl^{-1}.

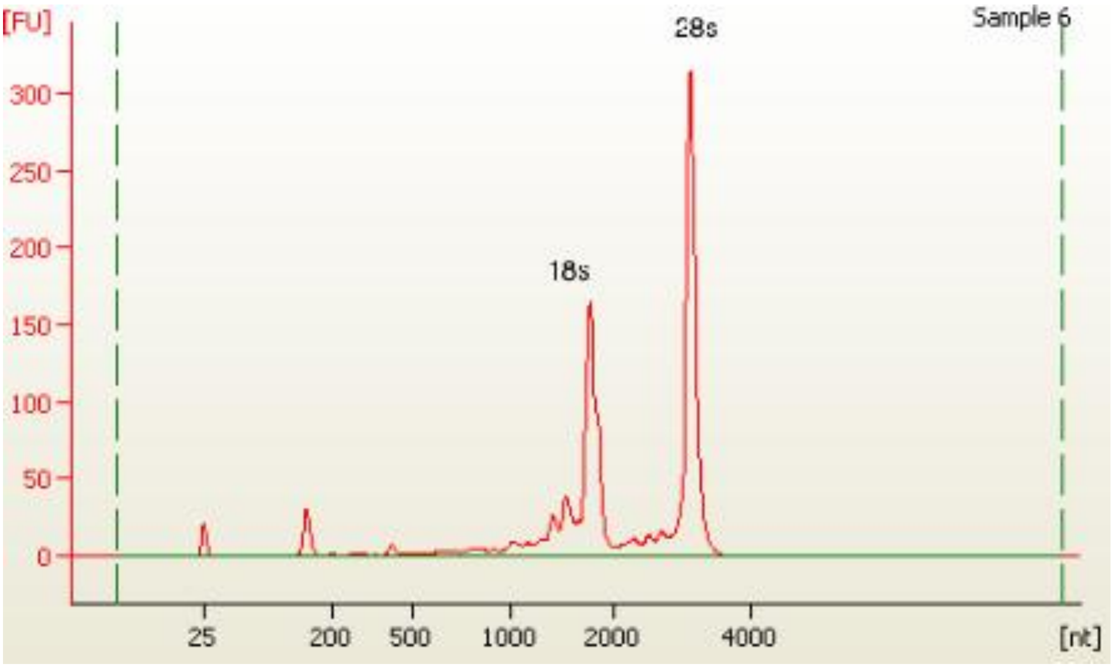


Figure 3-9 Total RNA electropherogram from Carlton sample 2

The rRNA ratio (28S/18S) was 1.75 with a RIN of 8.6 and a final concentration of 466.81 ng µl^{-1}.

3.4.2.2 cDNA library preparation

The four samples were used to prepare RNA-Seq libraries as described in 3.3.1.2. The resulting libraries were analyzed for quality using a Quibit fluorometer and an Agilent High Sensitivity DNA chip. The results can be seen in table 3.4.

Table 3-4 cDNA library final analysis.

The concentration was determined by the Quibit analysis and the average fragment size from the High Sensitivity DNA Chip.

Sample	**Concentration (µg µl^{-1})**	**Molarity (nM)**	**Average Fragment size (bp)**
Andrew's Choice 1	0.24	1.04	355
Andrew's Choice 2	0.26	0.86	468
Carlton 1	0.68	3.2	328
Carlton 2	0.90	3.4	405

Electropherograms of the resulting libraries can be seen in Figures 3.10 to 3.13. Ideally for Illumina sequencing a molarity of over 1nM is required and therefore Andrew's Choice sample 1 was selected for sequencing. Of the two Carlton samples it was decided that sample 2 should be used for sequencing as it has a higher molarity and higher average bp size of fragments.

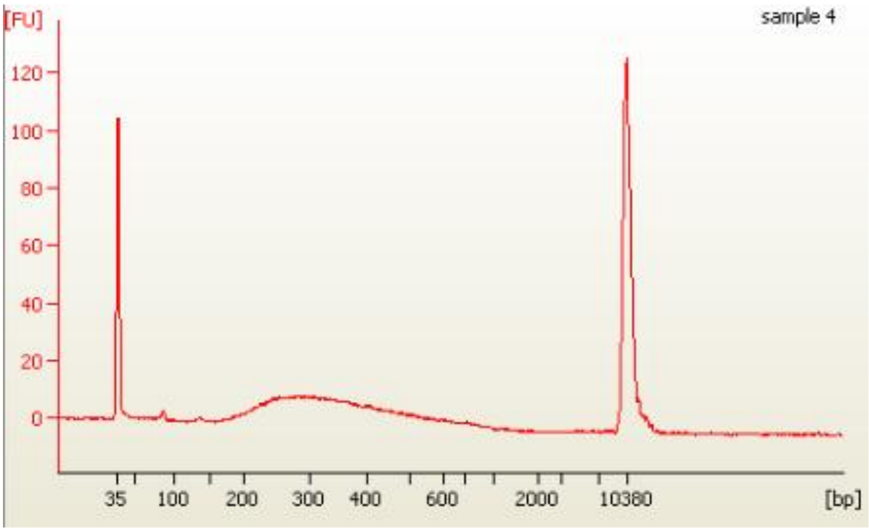


Figure 3-10 cDNA library electropherogram for Andrew's Choice sample 1.

The marker peaks are clearly visible at 35 and 10380 bp. The library is evenly fragmented, as no large peaks are visible. Ideally the fragmented area should show more distinct peaks but as this is a valuable sample it will still be sequenced as the molarity is above 1nM and the average fragment size is 355.

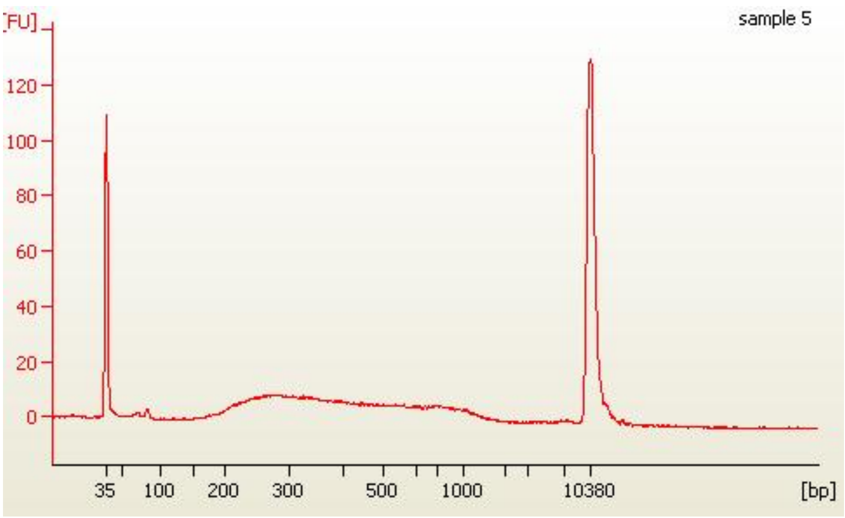


Figure 3-11 cDNA library electropherogram for Andrew's Choice sample 2.

The marker peaks are clear at 35 and 10380 bp. The fragmented area is more widely dispersed and of lower molarity than that of sample 1 and therefore will not be used for sequencing.

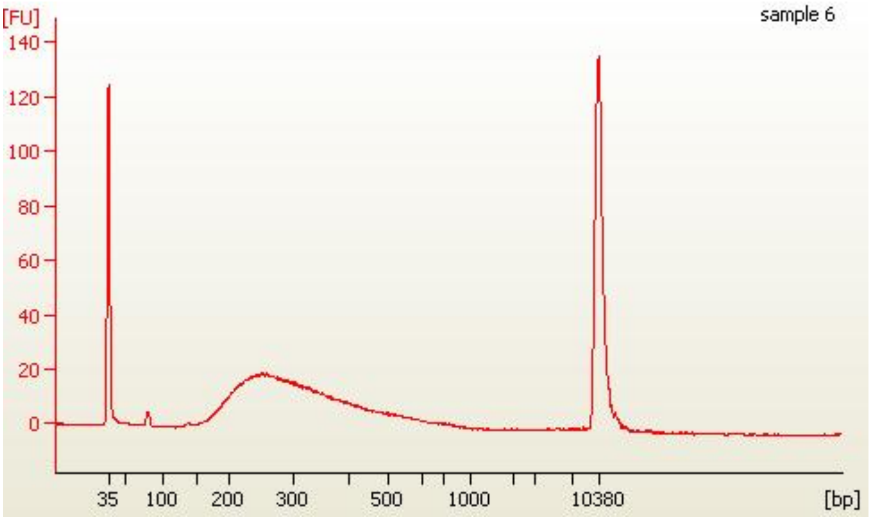


Figure 3-12 cDNA library electropherogram for Carlton sample 1.

The marker peaks are clear at 35 and 10380 bp. The fragmented region is of a much higher molarity than that seen for Andrew's Choice, however the average fragment size for this library was the lowest of the four, 328bp.

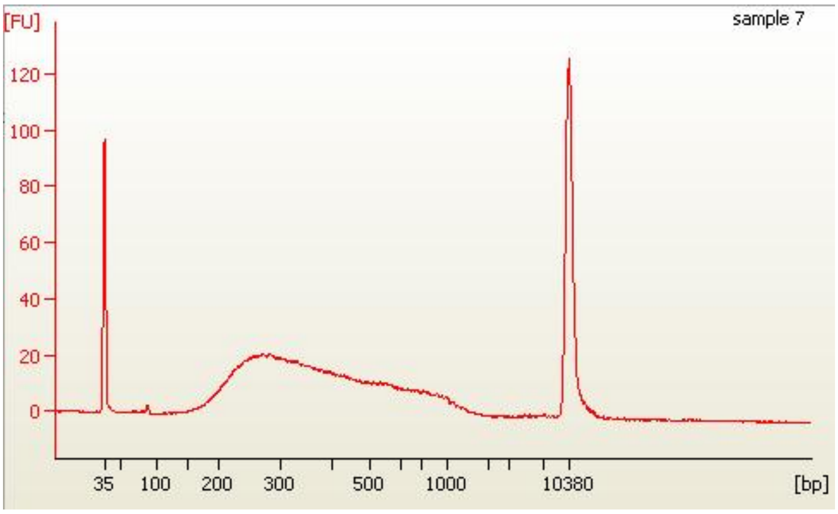


Figure 3-13 cDNA library electropherogram for Carlton sample 2.

The marker peaks are clearly visible at 35 and 10380 bp. The fragmented region is comparable in molarity and spread to that of Carlton sample 1. However the average fragment size is much larger, 405bp, therefore this sample was selected for sequencing.

3.4.2.3 Trimming and Filtering of raw reads

The results of the adaptor trimming and quality filtering and trimming can be seen in tables 3.5 and 3.6. Both varieties showed high percentages of high quality reads after the final trimming step (96.6% forward and 95.4% reverse reads for Carlton and 96.56% forward and 95.21% reverse reads for Andrew's Choice) with clearly reduced levels of non-ATCG bases (0.01% forward and 0.11% reverse reads for both varieties). The reverse reads from both varieties showed higher levels of non-ATGC bases and therefore slightly lower levels of high quality reads.

Table 3-5 Comparison of Carlton Illumina reads after adaptor trimming, quality filtering and quality trimming.

The percentage of reads with non-ATGC bases is significantly reduced without significantly reducing the number of high quality reads.

Sample name	Carlton forward post Cutadapt	Carlton reverse post Cutadapt	Carlton forward post filtering	Carlton reverse post filtering	Carlton forward post trimming	Carlton reverse post trimming
Minimum read length	1	1	1	1	20	20
Maximum read length	100	100	100	100	100	100
Average read length	94.52	94.68	94.42	94.48	94.17	94.05
Total number of reads	15575323	15575323	14825941	14825941	14806569	14806569
Total number of reads with non-ATGC bases	246413	1234958	206265	1101884	189590	1063571
Percentage of reads with non-ATGC bases	1.58%	7.93%	1.39%	7.43%	1.28%	7.18%
Total number of bases	1472165118	1474613644	1399937253	1400746087	1394344344	1392555397
Total number of high quality bases	1443938440	1410384386	1383593690	1377280745	1346885430	1328459060
Percentage of high quality bases	98.08%	95.64%	98.83%	98.32%	96.60%	95.40%
Total number of non-ATGC bases	291377	5274180	235489	1600297	205235	1556979
Percentage of non-ATGC bases	0.02%	0.36%	0.02%	0.11%	0.01%	0.11%

Table 3-6 Comparison of Andrew's Choice Illumina reads after adaptor trimming, quality filtering and quality trimming.

The results are very similar to those of Carlton suggesting that both varieties were sequenced equally effectively.

Sample name	Andrew’s choice forward post Cutadapt	Andrew’s choice reverse post Cutadapt	Andrew's choice forward post filtering	Andrew's choice reverse post filtering	Andrew's choice forward post trimming	Andrew's choice reverse post trimming
Minimum read length	1	1	1	1	20	20
Maximum read length	100	100	100	100	100	100
Average read length	93.19	93.43	93.11	93.15	94.99	94.86
Total number of reads	13945808	13945808	13187964	13187964	12853764	12853764
Total number of reads with non-ATGC bases	215968	1098861	178954	973819	164889	943026
Percentage of reads with non-ATGC bases	1.55%	7.88%	1.36%	7.38%	1.28%	7.34%
Total number of bases	1299547169	1302946984	1227939145	1228479921	1221023045	1219308786
Total number of HQ bases	1273925396	1237555422	1213408193	1207208031	1179003144	1160861352
Percentage of HQ bases	98.03%	94.98%	98.82%	98.27%	96.56%	95.21%
Total number of non-ATGC bases	255426	4638930	204560	1408963	178537	1373991
Percentage of non-ATGC bases	0.02%	0.36%	0.02%	0.11%	0.01%	0.11%

3.4.2.4 *Comparison of de novo assembly using Trinity and SOAPdenovo-Trans*

The results of both assemblies are shown in table 3.7. Trinity produced much longer contigs than SOAPdenovo-Trans. The latter assembly's shortest length is the same as the maximum length of the raw reads. The average read length (183bp) in SOAPdenovo-Trans is also low compared to the Trinity 511bp. The longest length seen by Trinity (30656 bp) is unlikely to be a true transcript due to the difficulty of assembling short reads into transcripts of this size and could be caused by repetitive sequences or contamination. (See section 3.4.2.10)

Table 3-7 Comparison of *de novo* assemblers.

Assembler	Trinity	SOAPdenovo-Trans
Number of contigs and singletons	165905	17429 (3458 contigs and 13971 singletons)
Average length (bp)	511	183
Shortest length (bp)	201	100
Longest length	30656	6864
N50	679	236

3.4.2.5 *Mapping of trimmed Illumina reads to the Trinity, SOAPdenovo-Trans and 454 assemblies*

The results can be seen in table 3.8. It is clear that for both varieties the 454 reference and SOAPdenovo-Trans assembly had much higher depth of coverage and lower number of total transcripts. By mapping the Illumina reads back to both the 454 and the SOAPdenovo-Trans reference it was clear that a large proportion of the reads were not mapping, suggesting that both references did not cover the entire transcriptome (only about 30% of the Illumina reads were accounted for in this mapping back to both the 454 reference and SOAPdenovo-Trans assembly (See section 3.5.2). The Trinity assembly gave better overall representation of the transcriptome.

Table 3-8 Comparison of mapping to the 454, Trinity and SOAPdenovo-Trans assemblies.

Variety	**Carlton**			**Andrew's Choice**		
Carlton Reference used to map to	**454 reference**	**Trinity assembly**	**SOAPdenovo-Trans assembly**	**454 reference**	**Trinity assembly**	**SOAPdenovo-Trans assembly**
Percentage of reference hit by Illumina reads	86.45%	95.42%	99.72%	82.85%	73.42%	97.24%
Percentage of Illumina reads that could be mapped back to the reference	30.84%	75.65%	39.91%	26.02%	56.38%	24.58%
Depth of coverage	72 X	12X	111X	61X	12 X	92X

The 454 reference has much higher coverage and the percentage of the reference hit is more similar for the two varieties. All three assemblies involve only Carlton reads and so a lower level of mapping would be expected for the Andrew's Choice reads, however the high levels of mapping seen with Andrew's choice agree with the idea that the two varieties are closely related.

3.4.2.6 SNP calling using custom SNP caller and VarScan

The full VarScan and pileup_parser.pl outputs can be found on the appendix disc (pileup_parser is in the scripts folder on the appendix disc). A basic comparison of the number of SNPs found in both assemblies can be seen in table 3.8.

Only the varietal differences are shown. For full inter and intra varietal SNP data see table 3.15 in the SNP calling in both the original and TE removed assembly section of this chapter.

A more in-depth study into these SNPs and the transcripts is described in chapter four.

Table 3-8 Intervarietal SNPs called using the two references.

Reference mapped	No of SNPs discovered
454	4032
Trinity	5766

The parsing script pulls out every alternative allele but has no constraints. VarScan has stringent cut offs but only shows one alternative allele. Combining the two methods results in the discovery of the true allelic variation in triploid species.

3.4.2.7 The prediction of ploidy level in Carlton and Andrew's Choice

The pileup_parser.pl results were parsed and the percentage of loci that showed more than one nucleotide at a specific loci were analysed. Out of the 23949 loci predicted to differ from the reference 98.09% were labeled "triploid" for Carlton and 98.54% were labeled "triploid" for Andrew's Choice. To be considered triploid the alternative alleles and the reference allele had to have an overall read percentage of 20 % or higher with the forth allele having a percentage of less than 10%. The remaining ~2% in both were labeled as tetraploids. To be considered tetraploid all four alleles had to be represented by 20% or more of the total number of reads.

3.4.2.8 Transcript level differences determined using BitSeq

Table 3.10 shows the number of contigs or singletons that are predicted to have significant transcript level differences between the two varieties and the percentage of the total number of transcripts this represents. The 454 data set resulted in 0.25% with a PPLR >0.95 (suggest a significant probability that the transcripts are up-regulated in the Andrew's Choice individual) and 1.34% with a PPLR <0.05 (suggesting down-regulation in the Andrew's Choice individual). The Illumina Trinity data resulted in 0.23% >0.95 and 0.29% <0.05. These transcripts are investigated in chapter four. The data shown is a summary of the results, showing only the PPLR values, the full BitSeq output containing PPLR, mean log2 fold change, confidence intervals and mean condition and mean expressions are shown in on the appendix disc. Only PPLR values are shown in table 3.10 as BitSeq suggests that these values can be used to “rank transcripts based on DE belief”. Although PPLR and p values are not interchangeable PPLR is still considered a viable method for predicting differences and can be used like p-values with a user defined cutoff.

Table 3-9 BitSeq results for both the 454 and Trinity assemblies.

Showing the difference in transcript levels between two individuals from two varieties of *Narcissus pseudonarcissus* (Carlton and Andrew’s Choice).

Assembly	**454**	**Trinity**
Total number of contigs/singletons	20349	165905
Transcripts with PPLR of >0.95	50	378
Transcripts with PPLR of <0.05	272	475
Percentage >0.95	0.25	0.23
Percentage <0.05	1.34	0.29

3.4.2.9 Annotation of Trinity assembly

The annotation pipeline described in chapter 2.3.4.5 was used on the Trinity assembly. It resulted in the annotation of 63826 transcripts (38.47%). The four databases searched and the number of unique hits found in each can be found in table 3.17 compared to the annotation of the post TE removal assembly.

3.4.2.10 Removal of transposon elements from raw reads – comparison of three methods

Table 3-10 Comparison of three methods of TE discovery.

The reads that hit TEs were removed from the raw and trimmed and filtered reads using all three methods to compare the resulting assemblies.

Method	**Variety**	**Number of reads removed**		**Total number of reads**		**Percentage of reads removed**	
		Raw reads	**Trimmed and filtered reads**	**Raw reads**	**Trimmed and filtered reads**	**Raw reads**	**Trimmed and filtered reads**
BMTagger	**Carlton**	39446	41133	15575323	14806569	0.253	0.278
	Andrew’	4474	54113	1394580	1285376	0.321	0.421

	s choice	5		8	4		
BLAST	**Carlton**	1428	1158	15575323	14806569	0.009	0.008
	Andrew's choice	1174	917	13945808	12853764	0.008	0.007
TransposonPSI	**Carlton**	60788	Not ran	15575323	14806569	0.390	Na
	Andrew's choice	36790	Not ran	13945808	12853764	0.264	Na

It is clear from table 3.12 that the BLAST search against TREP gave the lowest number of predicted TE transcripts; both BMTagger and TransposonPSI resulted in greater but similar numbers. The main difference between these two methods was the run time. BMTagger runs in a matter of hours whereas TransposonPSI took several weeks to run. All three methods were carried forward to assembly for further analysis.

3.4.2.11Assembly comparison of the three TE removal methods

Table 3-11 Comparison of Trinity assemblies of the three TE removal methods.

Statistics were determined using the NGStoolkit statistical program N50stats.pl

	BLAST		**BMTagger**		**TransposonPSI**
	Raw reads	**Trimmed and filtered reads**	**Raw reads**	**Trimmed and filtered reads**	**Raw reads**
Total sequences	169689	165176	169518	165065	169191
Total bases	84360430	82035348	84014952	81805100	84698744
Min sequence length	201	201	201	201	201
Max sequence length	21617	30656	21617	30656	30656
Average sequence length	497.15	496.65	495.61	495.59	500.61
Median sequence length	294	294	294	294	295
N25 length	1526	1513	1511	1511	1543
N50 length	634	630	629	626	644
N75 length	301	300	300	300	302
N90 length	231	231	231	231	231
N95 length	215	215	214	215	215

As can be seen in table 3.13 all three methods gave very similar assemblies. The filtering and trimming step removed ~4000 sequences in each method. Therefore it was decided that filtering and trimming should be carried out prior to TE removal. As the BLAST and BMTagger produced a similar number of sequences and the same max sequence length, it was decided to carry the BMTagger removal forward since it was quicker than BLAST and removed the most raw reads. TransposonPSI was deemed too time consuming compared to the other methods with no clear advantage.

3.4.2.12 Mapping of TE removed reads to both the 454 reference and the TE removed Trinity assembly

The differences between the above results and those seen in table 3.14 are minimal, the results suggest that the removal of TE only improves mapping by <1% for all varieties and references. This is consistent with the <1% of raw reads removed via these methods.

Table 3-12 Comparison of raw reads mapped back to the 454 reference and the Trinity Illumina *de novo* assembly.

Percentage of reference hit is determined using the coverageStatsSplitByChr_v2.pl script to determine the percentage of the reference hit by the raw reads. Percentage of Illumina reads mapped is the alignment percentage given by Bowtie2 representing the percentage of raw reads that were successfully mapped to the reference. The depth of coverage is the average read coverage seen from the mapping.

Variety	**Carlton**		**Andrew's choice**	
Reference Assembly	**454 reference**	**Post-TE Trinity assembly**	**454 reference**	**Post-TE Trinity assembly**
Percentage of reference hit	86.49%	95.59%	82.78%	73.62%
Percentage of Illumina reads mapped	30.74%	75.27%	25.92%	56.31%
Depth of coverage	72 X	12 X	60 X	12 X

3.4.2.13 SNP calling in both the Original and Post-TE removal assembly

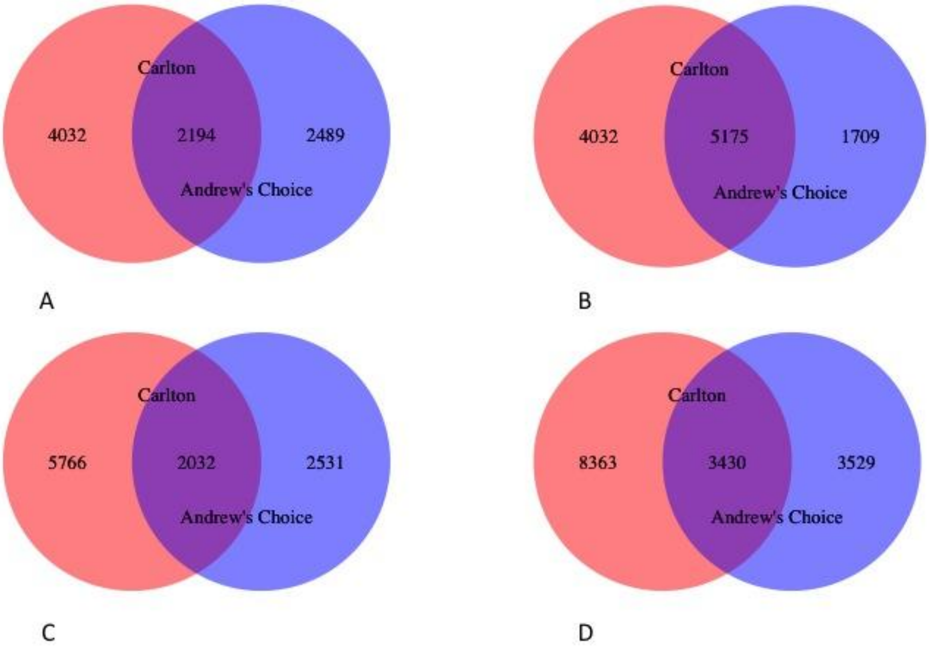


Figure 3-14 Venn Diagram showing SNP cross over between varieties

A shows the SNP cross over between Carlton and Andrew's choice in the original (no TE removal) 454 results. B shows the same cross over for the post TE removal results. C shows the original results for the Illumina data. D shows the cross over for the post TE removal Illumina results.

The inter-varietal SNPs range from 4032-8383 depending on the reference and if the raw reads have had the TEs removed or not. Transcripts that show the same intra-varietal and inter-varietal SNPs, and SNPs found in transcripts predicted to be involved in galanthamine production are discussed further in chapter four.

3.4.2.14 Transcript level differences

Table 3-15 BitSeq results for 454 reference, Original and Post-TE assemblies.

The 454 post-transposon removal represents the mapping of the filtered, trimmed and TE removed Illumina reads mapped to the initial 454 reference.

Reference	454	Original	454 post TE	Trinity post

		Trinity	removal	TE removal
Total number of transcripts	20349	165905	20349	165065
Transcripts with PPLR of >0.95	50	378	86	169
Transcripts with PPLR of <0.05	272	475	321	145
Percentage >0.95	0.25	0.23	0.42	0.1
Percentage <0.05	1.34	0.29	1.58	0.09

Table 3.15 shows the transcript levels in the two varieties are very similar with less than 1% showing PPLR of >0.95 and less than 2% showing PPLR of <0.05. The full BitSeq results including fold change differences can be seen in on the appendix disc.

3.4.2.15Annotation post TE removal

The Assembly from section 3.3.6 was annotated using the same pipeline as section 2.3.4.5. The pipeline annotated 62852 (38.1%) of the transcripts. The annotations from the four databases, TAIR, UNIPROT, Rfam and RefSeq were compared to those produced in 3.3.4 to evaluate any differences seen. The comparison is show in table 3.16.

Table 3-16 Comparison of assembly annotation of Original and Post-TE removal assemblies.

The annotation pipeline only allows each transcript to be annotated once, starting with TAIR. Those that are not annotated can then move onto the next database in the order of UniProt, Rfam and then RefSeq. The numbers shown represent the unique hits.

Database	**Number of transcripts annotated**		**Percentage of transcripts annotated**	
	Original Trinity assembly	**Post TE removal Trinity assembly**	**Original Trinity assembly**	**Post TE removal Trinity assembly**
TAIR	53827	52861	32.4	32
UniProt	2315	2270	1.4	1.4
Rfam	209	204	0.1	0.1
RefSeq	7475	7517	4.5	4.6
Total	63826	62852	38.5	38.1

3.5 Discussion

3.5.1 Production of cDNA libraries and read quality control

As discussed in section 2.1.7, extraction of RNA from plants with high levels of phenols, sugars and secondary metabolites is difficult. The CTAB method used in the production of the 454 reference was not suitable for extraction from Andrew's Choice. The Analytik-Jena InnuPREP plant RNA kit with the PL lysis buffer (specifically designed for high phenolic plants) was found to be a suitable method and was used for both varieties. From these samples four cDNA libraries were prepared and then one from each variety was carried forward to sequencing, sample 1 for Andrew's Choice as its molarity was over 1nM (recommended limit for Illumina) and sample 2 from Carlton as its molarity was 3.4 compared to sample 1 at 3.2nM. The sequencing of these samples resulted in two paired-end data sets with similar numbers of reads (average read lengths 100bp)(Andrew's Choice 13945808 pairs and Carlton 15575323 pairs). The raw reads were first trimmed using Cutadapt to remove the index primers and then filtered and trimmed to remove any reads shorter than 20bp, then those with less than 70% of bases having a quality score of 20 or more and finally bases from the end of reads that had a quality score of below 30. This resulted in the removal of ~5% of reads from each set with the reverse reads showing slightly higher levels of non ATGC bases. This could be due to several factors; base content can affect quality, also, if the forward and reverse reads do not overlap they can have different levels of quality and, finally, longer molecules can result in lower quality sequencing [180]. By trimming reads that had less than 70% bases with a quality score higher than 20 it was possible to produce high quality reads that could then be assembled with more confidence. As *de novo* assembly was required it was very important to remove bases and reads of low quality that could result in a non-representative assembly. It was also important to remove smaller reads (<20bp) that could again interfere with the assembly as both Trinity and SOAPdenovo-Trans work best on longer reads (~100bp).

3.5.2 Assembly of Illumina reads

Determining accuracy and coverage of an assembled transcriptome is much more difficult than that of a genome. It is further complicated if the transcriptome is from a non-model plant with no reference genome or closely related reference. There are several methods for assessing the quality of an assembly and they are discussed below in relationship to the daffodil data.

Initial results showed that Trinity produced much longer contigs, possibly suggesting a better assembly of the short reads. In transcriptomic *de novo* assemblies the use of N50 or the length of a transcript as a statistic for assessing an assembly is limited as unlike genomes where the longer the N50 the better, the assemblies of transcriptomes are intrinsically fragmented [132]. A higher N50 does not always relate to a better assembly and often those seen in transcriptome studies are higher than the actual N50 which could be linked to misassembly and concatenation of reads. However the short lengths of the SOAPdenovo-Trans assembly suggested a poor assembly, as the shortest length was the same as the maximum length of the raw reads. The average read length of 183bp from SOAPdenovo is also very low compared to average lengths in other transcriptomic projects. For example, the Trinity assembly result is much closer to the lengths seen in other Illumina assemblies such as *Pelargonium x hortorum,* (850bp Trinity) maritime pine (*Pinus pinaster*) (495bp) and *Centella asialica* (474bp) [111,127,181].

It is possible if a closely related species has been sequenced to produce a genome reference, or if there is a reference genome for the plant of interest that a Fermi estimation of the number of transcripts expected can be carried out [83]. Major plants have been predicted to have 20,000 to 40,000 genes, however not all of these will be present in the transcriptome and so an estimate can be made using micro array data along with a reference genome, for example it was predicted that the Arabidopsis leaf transcriptome would contain around 15,000 [83]. Polyploids and transcriptomes with a recent duplication event may have more but the overall number of contigs assembled can still be compared to other plant species. In the project by Xiao *et al* assembly from 75 species resulted in transcriptomes containing between 31000 and 70000 contigs [2]. While the diploid *Centella asiatica* transcriptome had 79041 tentative unique transcripts while both the diploid maritime pine (*P. pinaster*) and the allo-tetraploid *Nicotiana benthamiana* had much higher numbers of contigs (210513 and 235000 contigs) these are similar to those seen in Illumina.

One method of ascertaining the quality of an assembly is short read mapping, firstly the raw reads back to the assembly and also back to a reference genome if available [182]. The raw reads were mapped back to the 454, Trinity and SOAPdenovo-Trans assemblies. From the results it was clear that by using only the 454 assembly over 60% of the raw Illumina reads were not being exploited as only 30.84% (Carlton) and 26.02% (Andrew's Choice) of the Illumina reads mapped back to the 454 reference. Leaving out 60% of the Illumina reads would omit exploration of a significant data-set and so it was decided that a *de novo* assembly of the Illumina reads should be used alongside the 454 reference for further investigation such as SNP and transcript level analysis. Illumina sequencing has been shown to result in much deeper coverage of a transcriptome, as seen in this marked difference in coverage in this data set. A project by Zhang *et al* looking at two *Geraniaceae* transcriptomes showed that Illumina produced greater coverage [127]. This is not surprising as the Illumina sequencing produced almost 40 times the amount of data as that seen in 454, (similarly in this project the 454 produced around 150000 reads and the Illumina produced close to 15000000) however when they reduced the number of Illumina reads to match that of the 454 the resulting assemblies still showed a marked increase in coverage [127]. When the two techniques first appeared 454 had much longer read lengths and so was the platform of choice for *de novo* assembly, however as read length increased with Illumina and the improvement of *de novo* assemblers, Illumina is now more widely used. As is evident by the decommissioning of 454 [104].

No standard method currently exists for the combination of the two sequencing methods to produce one reference [2,181]. Of the two methods of Illumina *de novo* assembly, the SOAPdenovo-Trans reference had a much higher depth of coverage (111X and 92X compared to 12X for Trinity) for the section of the transcriptome it represented. However only ~30% of the raw reads were able to map back to the SOAPdenovo-Trans reference. This is similar to the percentage of raw reads that mapped back to the 454 assembly. The Trinity assembly resulted in a much higher use of the raw reads (75.65% and 56.38% compared to 26.02% and 24.58%) suggesting better overall representation of the sub-genomes of the transcriptome. This, and the low average transcript length seen with SOAPdenovo-Trans, resulted in the use of the Trinity assembly for further analysis as looking at any SNP differences between the homologous copies would be required to also assess the variation between the two varieties.

In plants with closely related reference genomes it is possible to compare the assembly to a reference and look at the transcripts ortholog hit ratio, that is the number of bases in a matched region divided by the length of the best matched sequence. An OHT of 1.0 suggested complete transcripts, however this assumes that the best match is indeed an ortholog with conserved length [182]. This is obviously linked to evolutionary distance, duplication events, loss of genes or silencing of one genome by another in polyploids can greatly affect this statistic as novel or significantly different transcripts may not map or may map poorly [182]. As no closely related genome exists for daffodils a more useful statistic may be the percentage of transcripts with unique hits to known plant proteins (specifically those known to be well conserved) this is discussed in section 3.5.4.

3.5.3 Removal of TEs

As *de novo* assembly of short reads is very difficult, longer contigs such as those seen in genome assemblies often suggest repetitive regions, TEs or concatenated transcripts. Although these repetitive regions are not thought to be transcribed the large maximum contig length of 30656bp in the Trinity assembly could be caused by concatenated transcripts, contamination or as suggested by a personal communication with the Trinity developers TEs. Therefore investigation into TEs was carried out on the Trinity data after the initial annotation step. This was originally carried out on the raw reads and then on the trimmed and filtered reads. Since these latter steps removed almost 5% of the raw reads, it would be beneficial to carry out filtering and trimming prior to the TE removal as leaving in this 5% of unfiltered reads could affect the assembly.

The method with the most stringent cut off was used to remove TEs as this would reduce the amount of data and should also reduce the number of false transcripts analysed. The BLAST search against TREP gave the lowest number of possible TE transcripts and so was not used for further investigation as the aim of this step was to remove as many erroneous reads as possible. The other two methods gave very similar results (0.2-0.4% reads removed) but took very different lengths of time (BMTagger took less than 2 hours whereas TransposonPSI took over a month). Therefore BMTagger was considered the most suitable method. Even then, the relatively low number of reads removed did not result in the removal of the longer transcripts. Although the use of basal plate was in part to avoid chloroplast contamination it may be that some of these longer transcripts are chloroplast associated. More over as there was no way of sterilizing the bulbs prior to RNA extraction it is possible that there was some bacterial contamination. It would be beneficial to a project of this nature to remove any transcripts that are linked to bacteria or chloroplasts if the overall aim was to produce a full-annotated transcriptome. However as only one tissue at one time point was used in the creation of the library the whole transcript is unlikely to be represented. The annotation pipeline did not annotate these longer transcripts, essentially excluding them from the putative gene search and due to time constraints and as the aim of the project was to look for putative genes involved in galanthamine production, no further analysis was carried out on the longer transcripts.

3.5.4 Annotation of Transcripts

To look for putative genes involved in alkaloid production the original 454 reference and Trinity assembly were used for transcriptome annotation, transcriptome wide SNP discovery and transcript level analysis and the results were compared to those gained for the data sets with the TE associated transcripts removed. Annotation was carried out on both Trinity assemblies in the same way as that for the 454 reference in chapter two. The results in table 3.17 show very little difference between the two assemblies with both resulting in ~38% annotation. This is comparable to other transcriptomic projects in plants; in the diploid *P. hortorum* and *Geranium maderense* (2n=68) Zhang *et al* reported an annotation rate of 37% and 49% using a number of assemblers including Trinity and SOAPdenovo-Trans [127]. Similar annotation levels were seen in a velvet assembly in *Centella asiatica*, a combined assembly of *Pinus pinaster* (using MIRA, CAP3 and ABySS) and a slightly higher average was seen in Xiao's 75 plant project (using MIRA for 454 and Velvet-OASES for Illumina reads) (53.04%, 46.6% and 68%) [2,111,181]. The exact breakdown of the annotations and further investigation such as Gene Ontology annotations to look for GO enrichment and functional information between the two assemblies and their use in the discovery of putative genes involved in galanthamine production is described in detail in chapter four.

3.5.5 SNP and transcript level difference analysis

Transcriptome wide SNP discovery and transcript level differences were also carried out on both data sets (Original and post TE removal). This was carried out to look for differences between the two varieties as they are closely related but there is very little to no data available on their genetic similarities. As they were grown under the same conditions and are known to produce different levels of galanthamine it was hoped that any difference seen in their transcriptomes could lead to the discovery of putative transcript differences linked to the biosynthesis of galanthamine. The results suggest that they have similar transcriptomes. Not only were Andrew's Choice raw reads mapped back to both the 454 and Trinity Carlton assemblies well (>70%) but their transcript levels show very few differences. In the 454 data the proportion of transcripts that showed significant differences between the two varieties were 0.25% in the original and 0.42% post TE removal for those with a PPLR above 0.95 (up-regulated in the Andrew's Choice individual) and 1.34% and 1.58% for those with a PPLR <0.05% (down-regulated in the Andrew's Choice individual). The change in number following the TE removal could be due to the improved mapping of quality raw reads to the 454 reference. The Illumina assembly also showed low levels of difference, 0.23% and 0.1% with PPLR >0.95 and 0.29% and 0.09% with PPLR <0.05%. The change in percentage seen here could be a direct result of TE removal producing a more accurate assembly for transcript level analysis. The next step in the analysis of these results is the investigation into transcripts that show different levels and the identification of any that could be involved in alkaloid production. Ideally for studies of this kind involving differential expression analysis the most accurate estimates are required, therefore it is standard practice and often a requirement of widely used differential expression methods to have replicates. In a study of this kind it is important to account for all sources of variation and so both technical and biological replicates are best practice [171]. Replicates can be used to normalize data and to account for changes in transcriptomic expression at different time points. However due to financial constraints repeating of the sequencing runs were not possible in this project. The data shown only compared one Carlton plant to one Andrew's choice plant at one time point, therefore this can only be used as a guide for further investigation. In order to look more closely at the differences between the varieties qPCR will be carried out involving both technical and biological replicates to look at transcripts that show significant differences in the sequence based differential expression analysis. This is described in chapter four.

SNP markers are a useful method of determining differences between varieties linked to phenotypes such as alkaloid production. As discussed in the introduction of this chapter (section 3.1.7.2) VarScan is capable of discovering SNPs in polyploids but only considers them true SNPs if the frequency of the minor allele is >0.5. Therefore by using VarScan to determine the loci of SNPs alongside the pileup_parser.pl script that gives raw values for each alternative allele it is possible to look for SNPs with frequencies below 0.5 as would be predicted in polyploids. The results of the transcriptome wide SNP discovery resulted in a range of inter-varietal SNPs (4032-8363) depending on the assembly used. The relatively low number of transcripts with inter-varietal polymorphisms compared to the total number of transcripts suggests that the two varieties are very similar; this is of particular interest when looking for differences in transcripts linked to secondary metabolites. Further investigation into inter-varietal SNPs in genes predicted to be involved in alkaloid production was the next step, as discussed in chapter four.

3.5.6 Bioinformatics approach to predicting ploidy level of Carlton and Andrew's Choice

The pileup_parser.pl script predicted ploidy level at individual loci to give a transcriptome wide view of ploidy within the sub-transcriptomes. It does not however take in to account any quality of reads and so sequence errors are not removed from the analysis. This script allows for further confirmation of ploidy alongside chromosome counting for an overall view. It also gives information on the heterozygosity of the sub-transcriptomes. Both Carlton and Andrew's Choice showed ~98% of loci with variation to be triploid. The majority of loci were the same as the reference with ~86000-88000 loci labeled "same" (the three nucleotides different from the reference having 0 reads) and ~1800000-2005000 labeled "nothing" (the reference nucleotide is represented by 25% or more of total reads with the three other alleles representing less than 15%). The SNP percentage was 1.2% in Carlton and 1.01% in Andrew's Choice which is comparable to a predicted SNP frequency of 1 per 100-300 bp, this is lower than that seen in the wheat and maize genomes (1 in 20bp and 1 in 70bp) [22]. A lower frequency would be expected in a transcriptome as only the transcribed region is analysed and so any SNPs present in introns or non-transcribed regions would be lost. The rate of variation seen in Andrew's Choice (1.01%) when mapped to the Carlton reference suggests that the two varieties are closely related, in agreement with the limited information available on the pedigree of Andrew's Choice (personal communication, Alzeim Ltd). This script could be developed further to incorporate quality and total read numbers as well as any available ratios on ploidy level at an individual level in allo or auto polyploids to predict ploidy level in any organism. This is discussed in Chapter 5 section 5.6.5.

4 Chapter Four: Analysis of transcriptome data to determine putative transcripts linked to Amaryllidaceae alkaloid biosynthesis

4.1 Introduction

4.1.1 Methods of annotation for gene function prediction in non-model *de novo* transcriptomes

The analysis in chapter three gave a generalized annotation based on similarity to sequences from a wide variety of databases. Assigning genes an annotation based on sequence similarity alone can lead to incorrect predictions of biological functions. Although secondary metabolite biosynthesis has been shown to involve only a relatively small number of gene families, these families contain large numbers of enzymatic isoforms (paralogs) that may or may not share the same or similar function. If these new paralogs have a different substrate specificity or altered kinetic characteristics they could also have a new biological role [183]. This emergence of paralogs is caused by gene duplication (one of the major sources of evolution) and has in turn generated the large number (>200 000) and diversity seen in plant secondary metabolites [183]. In order to identify genes involved in secondary metabolism it is therefore important to accurately predict the function of the genes as well as give an annotation based around sequence similarity. Numerous databases are now available that try to link the metabolite or gene to its biological function and these can be used to create an annotated backdrop for pathway analysis known as enrichment studies [184].

Three key databases that can be used for this purpose and that have been successful in other alkaloid biosynthesis studies are the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) and Enzyme Commission number (EC number) [2,185-187] and these are discussed below.

4.1.1.1 Gene Ontology (GO)

GO is one of the most successful examples of 'systematic description of biology' and is widely used in whole genome/transcriptome annotation projects [188]. The Gene Ontology consortium began as a collaboration between the model organism databases of Flybase, Mouse Genome Informatics (MGI) and *Saccharomyces* Genome Database (SGD) [185]. The goal of the project was to create a universal nomenclature to aid in the knowledge transfer of biological roles of genes or gene products recorded in studies to date. The ontologies were created utilizing several databases including SwissPROT (http://www.ebi.ac.uk/uniprot), Genbank (http://www.ncbi.nlm.nih.gov/genbank/), PIR (http://pir.georgetown.edu), MIPS (http://mips.helmholtz-muenchen.de/proj/ppi/), YPD and wormPD (https://portal.biobase-international.com/cgi-bin/portal/login.cgi), Pfam (http://pfam.xfam.org), SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) and ENZYME (http://enzyme.expasy.org). The Ontology system is split into three modules:

> **Biological Process:** Terms in this group include cell growth and maintenance as well as more specific terms such as translation. They represent the biological "objective" of the gene or gene product, accomplished by one or more molecular processes [185].
>
> **Molecular Process:** This includes terms such as enzyme or more specific terms such as adenylate cyclase. These describe the biochemical activity of the gene product but not where or when the activity occurs [185].
>
> **Cellular Component:** This final module describes the location of the gene product activity, such as ribosome or Golgi apparatus [185].

Genes are associated with numerous terms and any linked terms are also inferred, so a very specific term will also be intrinsically linked back to the most basic term available [188]. The annotations are both manually curated and computationally assigned based on current biological knowledge [189]. The evidence used ranges from experimental (the most reliable annotation method) to indirectly derived information from computational studies. Although the latter allows for the annotation of genes from non-model organisms lacking experimental evidence, it also increases the risk of false positives [188,190].

Computational annotations such as sequence similarity predictions of paralogs or orthologs can cause false positives as gene duplication and speciation can result in a change of function [188-190].

4.1.1.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG was developed in 1995 as part of the Human Genome Program of the Ministry of Education, Science, Sport and Culture in Japan [186]. This database links genomic data with higher order functional information. It is often used in investigation into biosynthetic pathways as it is one of the largest and most comprehensive databases linking metabolism to gene available [184]. Its two main databases are:

> **Gene Database:** catalog of genes and annotations from complete and partial genomes [186].
>
> **Pathway Database:** graphical representations of cellular processes, representing higher order functional information [186]. An example would be the phenylpropanoid pathway, which is the starting point for numerous larger pathways such as Amaryllidaceae alkaloid biosynthesis.

KEGG has been used in numerous studies on secondary metabolites. However, it is important to review the results of this or any other functional database carefully [2,5,111]. KEGG contains a generalized set of pathways that are conserved throughout the plant, microbial and animal world as well as more specific pathways that stem from these. As part of this, there are sets of broad terms such as 'metabolism pathway' that have little meaning when assigned function [184]. Further complications can arise if a pathway contains either very many or very few enzymes or intermediates, since such pathways can bias downstream analysis such as enrichment studies. A final point to consider when scrutinizing the results of database searches is that many pathways are species specific and it is possible that a pathway that does not exist in a certain species or organism can be predicted to be present or show enrichment in later enrichment studies [184].

4.1.1.3 The Enzyme Commission (EC number)

The Enzyme Commission number is a numerical classification system for enzymes based upon the chemical reactions that they catalyze [191]. The original list was created in 1992 and has been updated periodically ever since [192]. It is a unified classification system that may help with annotation of unknown transcripts or genes.

The use of KEGG, EC and GO annotation gives important information on the functionality of genes but without analysing the relationship between genes within the transcriptome. As primary and secondary metabolism involved a very complex network of genes and pathways, with very similar genes carrying out different roles, it is imperative that the overall relationship between the predicted genes and their relative abundance as well as metabolite levels are analysed. The databases discussed above offer a backdrop for enrichment studies, also known as pathway analysis studies, that examine gene to gene interactions and relationships [184].

4.1.1.4 Enrichment analysis

Enrichment studies involve looking for genes that are represented by significantly larger or smaller numbers of reads within a transcriptome by comparing a list of genes that could potentially be involved in the pathway of interest to a background (often the whole transcriptome) [193]. It requires the genes of interest to be associated with the metabolites of interest and annotated with as much information as possible such as chemical family, metabolic pathway and gene family [184]. It is important when using GO, KEGG, EC or any other annotation, that enrichment alone is not considered conclusive. To be confident that a higher proportion of genes annotated with a specific term exist among the genes of interest compared against the whole data set, it is important that the enrichment score is compared to the probability of occurrence by chance. For example, if within a gene set of 100 secondary metabolism genes, 5 genes are annotated as PAL but there are only 6 PAL genes in the whole data, set this must be taken into account [188].

With an increase in sequencing data from second and third-generation techniques, methods for annotation enrichment have also grown rapidly and there is no current "gold standard" method. From 2005 to 2010 the number of tools available increased from 14 to 68 [193]. These tools have been classified into 3 types:

Singular enrichment analysis (SEA): this method looks at a list of genes deemed interesting or important and one gene at a time. It therefore does not take relationships into account [193]. By only utilizing GO terms, its analytical abilities are limited as, although rapidly increasing, the number of GO terms is still limited and often too broad to infer functionality.

Gene set enrichment analysis (GSEA): Although this looks at all genes in a data set, it examines each independently, thus missing important relationship details [194]. In addition, it is not well suited to the daffodil dataset since GSEA requires a quantitative biological value for each gene, such as fold change and differential expression [194]. Although a DE study was carried out on the daffodil data, it originated from only two individuals and so would not have the reproducibility required for a GSEA method.

Modular Enrichment Analysis (MEA): This method looks for relationships between terms leading to a reduction in redundancy although several MEA tools rely on only one source of annotation such as GO [193]. However some, such as the Database for Annotation, Visualization and Integrated Discovery (DAVID), has functional classification tools that use a range of sources to determine enrichment [194].

4.1.2 DAVID enrichment analysis

David is a "module-centric approach for functional analysis of large gene sets" that highlights over-represented biological terms [195]. DAVID compares annotation profile similarities between genes and so gives broader functionality groupings than methods relying on only, for example, sequence similarity or gene family determination [195]. It is necessary to combine as many annotation methods and characterisations as possible to infer functionality linked to secondary metabolite production. DAVID's knowledgebase brings together clusters of over 40 publicly available functional annotation databases [196]. By linking all available identifiers for a given gene to a single DAVID ID the data is centralized, speeding up analysis and functional annotation [196].

The analysis is carried out using EASE (Expression Analysis Systematic Explorer) scores with all individual EASE scores for all members of a group counting towards the overall group score, not just those genes on a list of interest [195]. EASE scores are a modified version of a one-tailed Fischer exact probability test [197]. These calculate the probability of random sampling of genes in a population in relation to the whole data set. For example if 10 genes in a sample list of 100 were annotated as methyltransferases and the background list of 1000 contained 12, the test would estimate the probability of finding the methyltransferases in both the background and sample set and compare the probabilities [197]. The EASE scoring takes into account the distribution of gene types and so adjusts for rarer or more common genes using a jackknife method [197]. This is particularly useful in biological studies since the Fischer exact test would make a rare category with just one gene significant if that gene was present in the list of interest rather than a category with a large number of genes in the main population and relatively few in the interest list [197]. EASE scores offer a compromise between the two extremes making more biological relevant probability predictions. DAVID utilizes EASE scores to group together clusters of genes with similar functions and looks for patterns of enrichment for the clusters containing the genes of interest.

Therefore, in order to carry out enrichment analysis, it is first important to create a suitable background and list of genes that are of interest (known as the GOI list in this project). The background must be appropriate to the gene of interest list. The full transcriptome could be used but the analysis often requires GO terms or UniProt IDs and only those transcripts that can be annotated can be used. A large number of transcripts are often not annotated in non-model plants and this can cause bias in an enrichment study [188]. A compromise must be made when annotating the background list. Although using a closely related species may lead to more accurate annotations, if this species has few annotations it will produce few annotations for the species of interest and so a better-described species, such as *Arabidopsis thaliana* in plant research, would be more appropriate [189]. For the gene of interest list database searches such as KEGG, GO and EC can be used as a starting point to find genes in similar pathways or families [5,111,198]. However alongside this it is useful to look at the research literature on genes and enzymes that have been identified in the biosynthetic pathways in other alkaloid producing plants to limit the number of predicted genes [18,199].

4.1.3 Alkaloid biosynthesis studies

Three of the most studied pathways, for economic and pharmacological reasons, are those involved in the production of codeine and morphine in *Papaver somniferum*, capsaicin and other capsaicinoids in *Capsicum annuum* and compounds such as rosmarinic acid from Lamiaceae herbs [15,200,201]. These and many other secondary metabolite pathways involve only a small collection of gene families that are used to create and modify a key precursor sometimes referred to as a scaffold (see tables 4.1, 4.2 and 4.3 for details of enzymes linked to secondary metabolites in poppies, Capsicum and Lamiaceae)[202]. The generation of these precursors involves the use of primary metabolite substrates as building blocks for several precursors generated through a small set of enzymatic reactions [202]. These precursor, such as strictosidine for terpene indole alkaloids and norcoclaurine for BIAs, are then modified to create the large and diverse number of metabolites seen within these secondary metabolite families [110,203,204].

Modification of the key precursor can occur either via redox chemistry such as oxidation or through group transfers such as alkylation, acylation and glycosylation [202]. The enzymes involved in these modifications can often be related but with different substrate specificity, or can be part of a pathway involving several different enzyme classes [202].

The following discussion focuses in on two classes of enzymes that are involved in alkaloid production in several plant systems and could be suggested for the galanthamine pathway, namely P450 monooxidases and Pictet-Spenglerases [17,203,205,206]. These will be discussed in relation to their involvement in secondary metabolism and the pathways chosen for investigation to create a database of predicted gene homologs.

4.1.3.1 P450 subfamilies linked to secondary metabolism

Cytochrome P450s are a class of haem enzymes named for their maximum optical absorbance at 450 nm (in a reduced state complex with carbon monoxide) [207]. They are found in numerous plant organelles including the endoplasmic reticulum, mitochondria, plastids and Golgi bodies [208]. The first plant P450 was identified in 1969 in cotton and the first to be sequenced was CYP71A1 from avocado [209,210]. Amino acid similarity is used in the classification of P450s. If the similarity between two is over 40% they belong to the same family. Within a family, a similarity above 55% classifies the proteins within the same subfamily [208].

P450s are involved in many processes in both primary and secondary metabolism and are predicted to account for up to 1% of plant genome annotations. For example there are 246 P450s annotated in Arabidopsis, 356 in rice, 312 in poplar and 457 in grape [211]. Within secondary metabolism P450s catalyse a variety of monooxygenation and hydroxylation reactions. They also catalyse four types of unusual reactions, specifically methylenedeoxy bridge formation, phenol coupling, oxidative rearrangement of carbon skeletons and oxidative C-C bond cleavage [211]. The first two types of reaction are now discussed in greater detail due to their involvement in alkaloid biosynthesis [211,212].

4.1.3.2 *Methylenedeoxy-bridge formation*

This reaction involves the formation of a bridge via oxidative cyclisation of an ortho-hydroxymethoxy-substituted aromatic ring [211]. Within isoquinoline alkaloid biosynthesis these reactions are catalysed by P450s. These reactions are deemed unusual since they do not involve the stereotypical hydroxylation step of inserting an oxygen molecule [213].

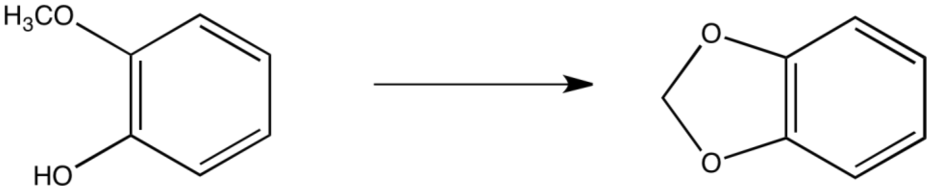


Figure 4-1 A simplified methylenedeoxy-bridge formation.

The clan responsible for bridge formation in isoquinoline alkaloid biosynthesis is CYP719A in a reaction where a hemiacetal intermediate of formaldehyde (formed via a P450-dependent hydroxylation of a methoxy group) is cyclized via an ionic mechanism to produce the methylenedeoxy-bridge [214-216]. These enzymes are all substrate specific and so several different CYP719As are responsible for similar reactions throughout alkaloid biosynthesis. CYP719A1 converts tetrahydrocolumbamine to (S)- tetrahydroberberine in canadine synthesis in Japanese goldthread (*Coptis trifolia*) [217]. In *Eschscholzia californica* two CYP719As are involved in stylopine synthesis but show differing substrate specificity. CPY719A2 has high affinity for (R, S)-cheilanthifoline alone, while CYP719A3 has affinity for three similar substrates, (R, S)-cheilanthifoline, (S)-scoulerine and (S)-tetrahydrocolumbamine [218].

4.1.3.3 *Phenol coupling*

The final step in the biosynthesis of galanthamine is an intramolecular para-ortho phenol coupling reaction. Reactions like this are seen in numerous alkaloid biosynthesis pathways and are often catalysed by CYP80 or CYP719 enzymes [205,212]. Intramolecular C-C phenol coupling is required in BIA synthesis and is catalysed by both CYP80G2 and CYP719B1 depending on the substrate. In *C. japonica* CYP80G2 converts (S)-reticuline to (S)-corytuberine whereas in the synthesis of salutaridine as part of the morphine biosynthetic pathway in *P.*

somniferum, CYP719B1 converts (R)-reticuline to salutaridine [205,219]. In contrast both the S and R configurations of N-methylcoclaurine in bisbenzylisoquinoline alkaloid biosynthesis can be converted by the same P450, CYP80A1, catalyzing an intermolecular C-O phenol coupling reaction to form berbamunine [220].

Both CYP719B1 and CYP80G2 show high levels of sequence similarity to other P450s but different substrate specificity and therefore catalyse different reactions (CYP719B1 shares 51% amino acid similarity with CYP719A1 and CYP80G2 shares 52% amino acid similarity with CYP80A1)[205,219]. It is therefore important to examine function as well as sequence similarity when predicting possible homologs in daffodils and this is particularly important in P450s. The fact that some P450s are highly substrate specific and others have broader substrate specificity makes it difficult to predict which family or subfamily will be involved in galanthamine biosynthesis and several will therefore be included in the predicted database.

4.1.3.4 PSRs role in secondary metabolism

Pictet Spengler reactions are condensation reactions employed in the synthesis of alkaloids, named after Amé Pictet and Theodor Spengler who discovered them [221]. They synthesised the alkaloid 1,2,3,4, tetrahydroisoquinoline via a cyclo-addition reaction between β-phenylethylamine and formaldehyde [221]. The first plant enzyme that catalysed this type of reaction was found in *Catharanthus roseus*. The Pictet Spenglerase (PSR) strictosidine synthase is involved in the biosynthesis of strictosidine, the key precursor of monoterpenoid indole alkaloids [222,223]. Strictosidine synthase was first purified in 1979 and since then has been used as a biomimetric synthase in the synthesis of novel alkaloids [206]. Although strictosidine synthase is a member of the 6-bladed β propeller protein family, not all PSRs are from this protein family [203,206,224].

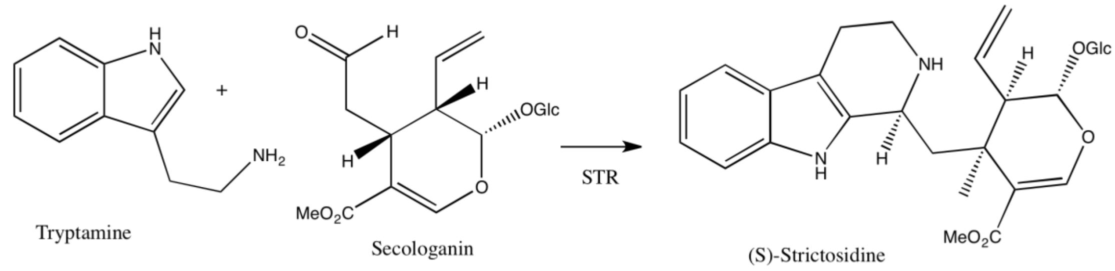


Figure 4-2 STR catalyzing a PSR reaction.(adapted from Stockigt *et al.*, 2011) [206].

PSRs are an example where several gene/enzyme families catalyse the same type of reaction. In BIA biosynthesis a similar reaction is carried out via a Bet v1/PR10 family protein known as norcoclaurine synthase (NCS) [204]. NCS catalyses an asymmetric PS condensation of dopamine and 4-hydroxyphenylacetaldehyde to synthesize (S)-norcoclaurine in the first dedicated step of morphine and codeine biosynthesis [18]. NCS has been isolated from several plants including *P. somniferum* and *Thalictrum flavum* [203,204]*.* Work has also been carried out to see if PSRs have similar ancestry or share sequence similarity. NCS and do not share any homology and so it is predicted that they evolved to carry out similar reactions from different ancestral proteins [225].

The study of both P450s and PSRs involved in secondary metabolism shows the need for analysis of sequence similarity, gene families, enzymatic reaction types and metabolite and transcript levels in the search for putative genes. In the search for possible genes involved in galanthamine production, all of the above methods will be utilized.

4.1.4 Well-studied alkaloid biosynthetic pathways in plants

4.1.4.1 Benzylisoquinoline alkaloids in Papaver somniferum

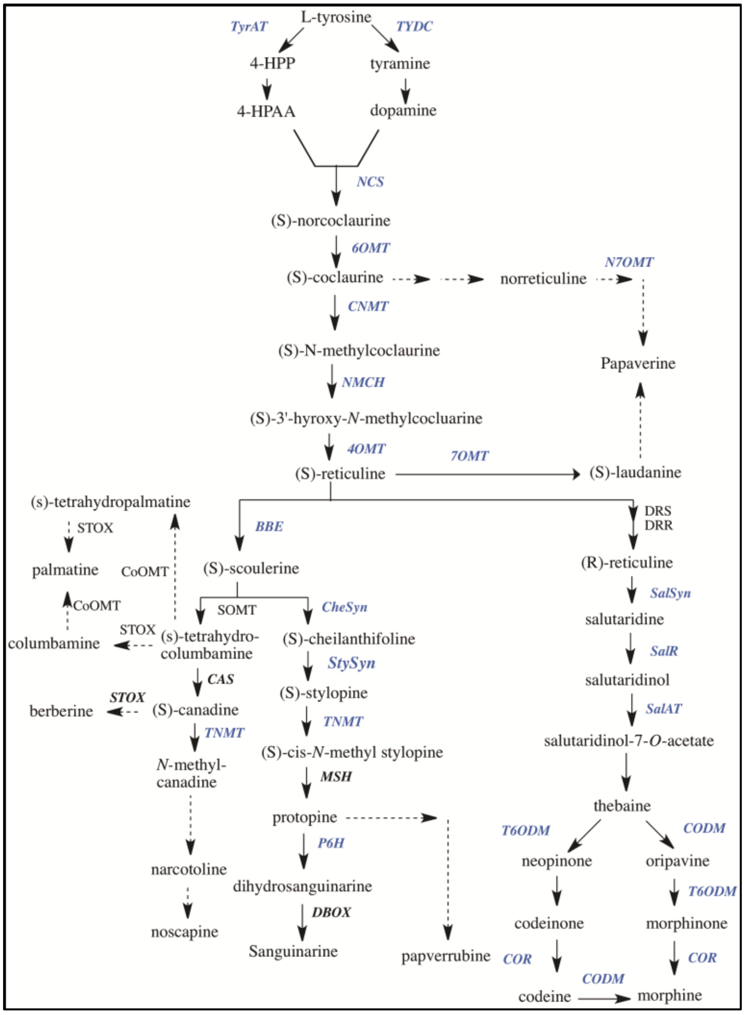


Figure 4-3 Biosynthesis of several BIA subtypes (Modified from Desgagne 2012).

The enzymes that as of 2012 had been isolated or characterized from BIA producing plants are shown in blue. This pathway was used to predict similar steps in the proposed galanthamine pathway. The enzymes that have been isolated or characterized are shown in table 4.1.

The benzylisoquinoline alkaloid (BIA) pathway of the opium poppy has been studied extensively and therefore the majority of the enzymes, or at least those involved in the production of morphine and codeine, have been characterized. Figure 4.3 shows the pathway with characterized enzymes in blue. Despite the variety seen within benzylisoquinoline alkaloids only a small set of protein families are responsible for their biosynthesis, namely cytochrome P450s, S-adenosylmethionine- dependent *O*- and *N*- methyltransferases, four distinct groups of NADPH dependent dehydrogenases/reductases, FAD-linked oxidoreductases, certain acetyl-CoA-dependent *O*-acetyltransferases, 2-oxoglutate/Fe(II)-dependent dioxygenases and carboxylesterases [8].

Table 4-1 The enzymes involved in the biosynthesis of codeine, morphine and sanguinarine that have been isolated or characterized.

Enzyme abbreviation	Full name	Source
TYDC	Tyrosine/dopa decarboxylase	[226]
TyrAT	Tyrosine aminotransferase	[114]
NCS	Norcoclaurine synthase	[18,204,227]
6OMT	(S)-norcoclaurine 6-O-methyltransferase	[228,229]
CNMT	(S)-Coclaurine N-methyltransferase	[230]
NMCH	(S)-N-methylcoclaurine 3'-hydroxylase	[231,232]
4OMT	(S)-3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase	[229]
N7OMT	Norreticuline 7-O-methyltransferase	[233]
7OMT	Reticuline 7-O-methyltransferase	[228]
SalSyn	Salutaridine synthase	[205]
SalR	Salutaridine reductase	[234]
SalAT	Salutaridinol 7-O-acetyltransferase	[235,236]
T60DM	Thebaine 6-O-demethylase	[113]
COR	Codeinone reductase	[237]
CODM	Codeine O-demethylase	[237]
BBE	Berberine bridge enzyme	[238-241]
CheSyn	Cheilanthifoline synthase	[216,242]
StySyn	Stylopine synthase	[218,242]
TNMT	Tetrahydroprotoberberine N-methyltransferase	[243]
P6H	Protopine 6-hydroxylase	[244]

4.1.4.2 Capsaicinoids in Capsicum annuum

The current state of research on this biosynthetic pathway is not as extensive as in poppies, but the biosynthetic pathway has been established and is shown in figure 4.4.

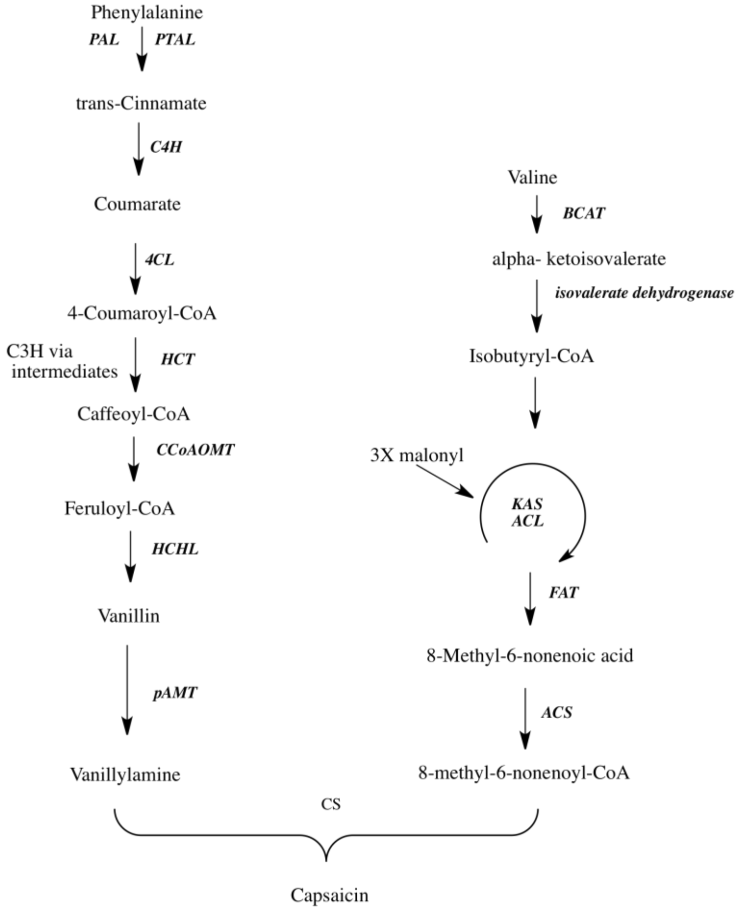


Figure 4-4 Biosynthesis of capsaicin (Modified from Aza Gonzalez *et al.*, 2010)

The enzymes involved in the pathway have not been studied extensively [200]. Capsaicinoids are derived from two precursors synthesised in separate pathways that come together at the point when a molecule of vanillylamine (from phenylalanine in a phenylpropanoid pathway) is condensed to a short branched unsaturated fatty acid [200]. The main enzymes involved in the production of capsaicin are shown in table 4.2 with a source for the discovery, isolation or prediction of them in *Capsicum*.

Table 4-2 The enzymes involved in the biosynthesis of capsaicin.

The full name and relevant references related to their study in varieties of pepper is also given.

Enzyme abbreviation	Full name	Source
PAL	Phenylalanine ammonia lyase	21,245,246
C4H	Cinnamate 4-hydroxylase	21,245,246
4CL	4-coumaroyl-CoA ligase	247
HCT	Hydroxycinnamoyl transferase	247,248
C3H	Coumaroyl shikimate/quinate 3-hydroxylase	245,246
CCoAOMT	Caffeoyl-CoA 3-*O*-methyltransferase	247
HCHL	Hydroxycinnamoyl-CoA hydratase/lyase	247,248
pAMT	Putative aminotransferase	21
BCAT	Branched-chain amino acid transferase	
KAS	Ketoacyl-ACP synthase	21,249
ACL	Acyl carrier protein	249
FAT	Acyl-ACP thioesterase	249
ACS	Acyl-CoA synthetase	250
CS	Capsaicin or capsaicinoid synthase	248,251,252

Study of this pathway may lead to a greater understanding of certain steps within the galanthamine pathway since PAL is known to be the first enzyme involved in both pathways [4,200]. The conversion of phenylalanine to protocatechuic acid is carried out via an acidic intermediate and is achieved via several steps involving enzymes including PAL (see figure 4.6 for a predicted pathway) [4]. 3,4-dihydroxybenzaldehyde (protocatechuic aldehyde) is structurally similar to vanillin, an intermediate in the phenylpropanoid pathway of capsaicin biosynthesis, as is show in figure 4.5.

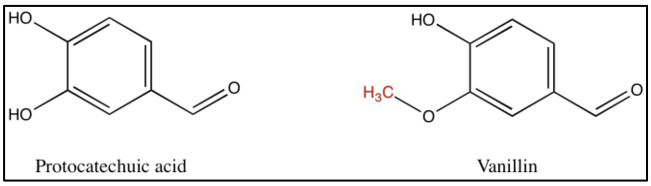


Figure 4-5 Chemical structure of protocatechuic acid and vanillin intermediates in the production of galanthamine and capsaicin.

The methyl group of the vanillin is highlighted to show that it is simply a methylated protocatechuic acid.

The only difference is a methyl group in vanillin, as highlighted in figure 4.5. A predicted pathway for the production of protocatechuic acid could be based on the production of vanillin. Biosynthesis could follow a similar path until the production of caffeoyl-CoA, with omission of the COMT/CCoAOMT catalyzed step resulting in a non-methylated structure. A prediction of this process is shown in figure 4.6.

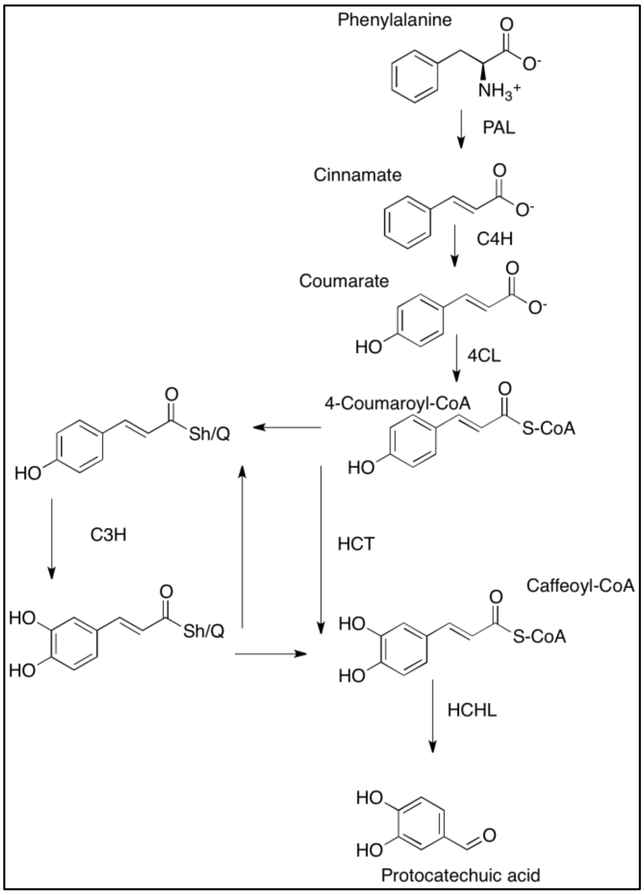


Figure 4-6 Possible pathway for the production of protocatechuic acid in galanthamine biosynthesis.

The pathway was predicted using the production of vanillin in capsaicin biosynthesis as a guide [200].

4.1.4.3 Rosmarinic acid (RA) in Lamiaceae herbs

Although rosmarinic acid is not an alkaloid but an ester, its biosynthesis involves several of the same steps or enzymes seen in poppies and peppers [15,200,253]. It is synthesized from a product of the phenylpropanoid pathway (caffeic acid) and 4-dihydoxyphenyllactic acid (HPLA), derived from tyrosine- [253].

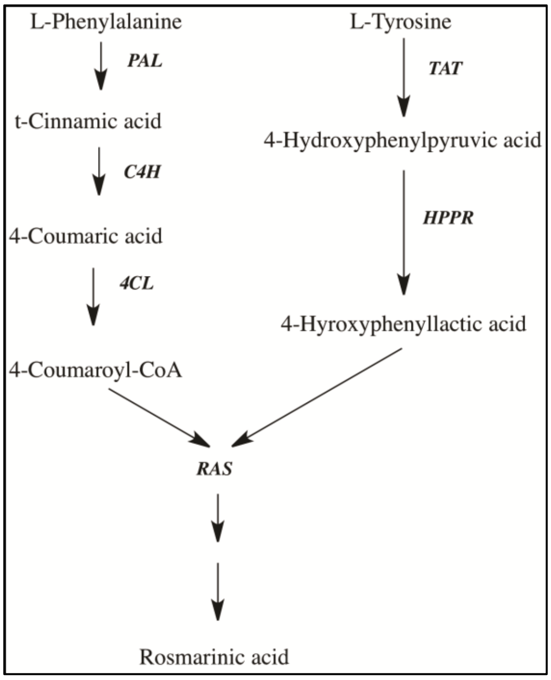


Figure 4-7 Biosynthesis of rosmarinic acid in Lamiaceae species (Modified from Xiao 2011)[20].

The pathway can be seen in figure 4.7 and the enzymes involved with associated references can be found in table 4.3. Not unlike the capsaicin pathway, RA biosynthesis involves the phenylalanine pathway, and offers homologs of the same enzymes seen in Capsicum that may or may not share more sequence similarity to putative genes in daffodils.

Table 4-3 The enzymes involved in rosmarinic acid biosynthesis.

Relevant sources are cited that discuss the study of these enzymes with RA biosynthesis.

Enzyme abbreviation	**Full name**	**Source**
PAL	Phenylalanine ammonia lyase	[213]
C4H	Cinnamate 4-hydroxylase	[254]
4CL	4-coumaroyl-CoA ligase	[255]
TAT	Tyrosine aminotransferase	[253]
HPPR	Hydroxyphenylpyruvate reductase	[20]
RAS	Rosmarinic acid synthase	[256]

4.1.5 Investigating the galanthamine pathway

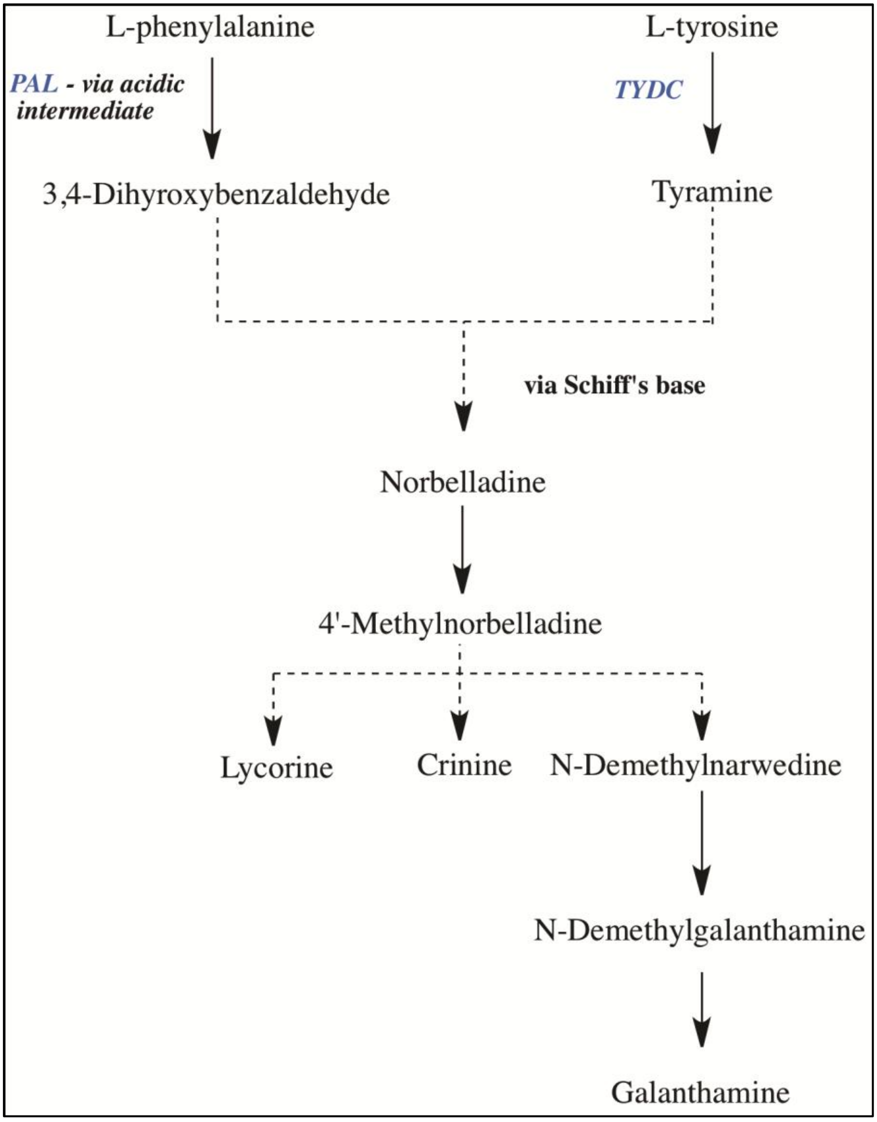


Figure 4-8 Proposed biosynthetic pathway of galanthamine.

(Modified from the isoquinoline alkaloid biosynthesis pathway of KEGG (http://www.kegg.jp/kegg-bin/show_pathway?rn00950+R08441). TYDC and PAL have been shown to catalyze the first steps of the reaction prior to this study and so are shown on the pathway in blue [199,257]. The reactions with solid lines have been experimentally proven; those with dotted lines are predicted steps.

The pathway (see figure 4.8) is understood to begin with L-phenylalanine and L-tyrosine and the first steps are known to involve the enzymes PAL and TYDC resulting in cinnamic acid and ammonia from the PAL pathway and tyramine from the TYDC route [199,257]. The cinnamic acid is then degraded to give protocatechuic aldehyde (3,4-dihydroxybenzaldehyde) and this, along with tyramine, is condensed via a Schiff base condensation reaction to give norbelladine, that is methylated to the precursor 4-*O*-methylnorbelladine [4].

From this precursor the variety of Amaryllidaceae alkaloid types are synthesized via three methods of C-C phenol coupling. The para-ortho coupling results in galanthamine-like alkaloids, ortho-para coupling in the lycorine like alkaloids and the para-para coupling results in crinine like alkaloids [258]. The enzymes that catalyze these reactions are currently unknown.

The aim of this chapter is to bring together the whole transcriptome annotation, SNP and transcript level profiles from chapter three with information on alkaloid/secondary metabolite production in other, better studied plant systems to produce a list of transcripts putatively involved in galanthamine biosynthesis in *N. pseudonarcissus*.

4.2 Methodology

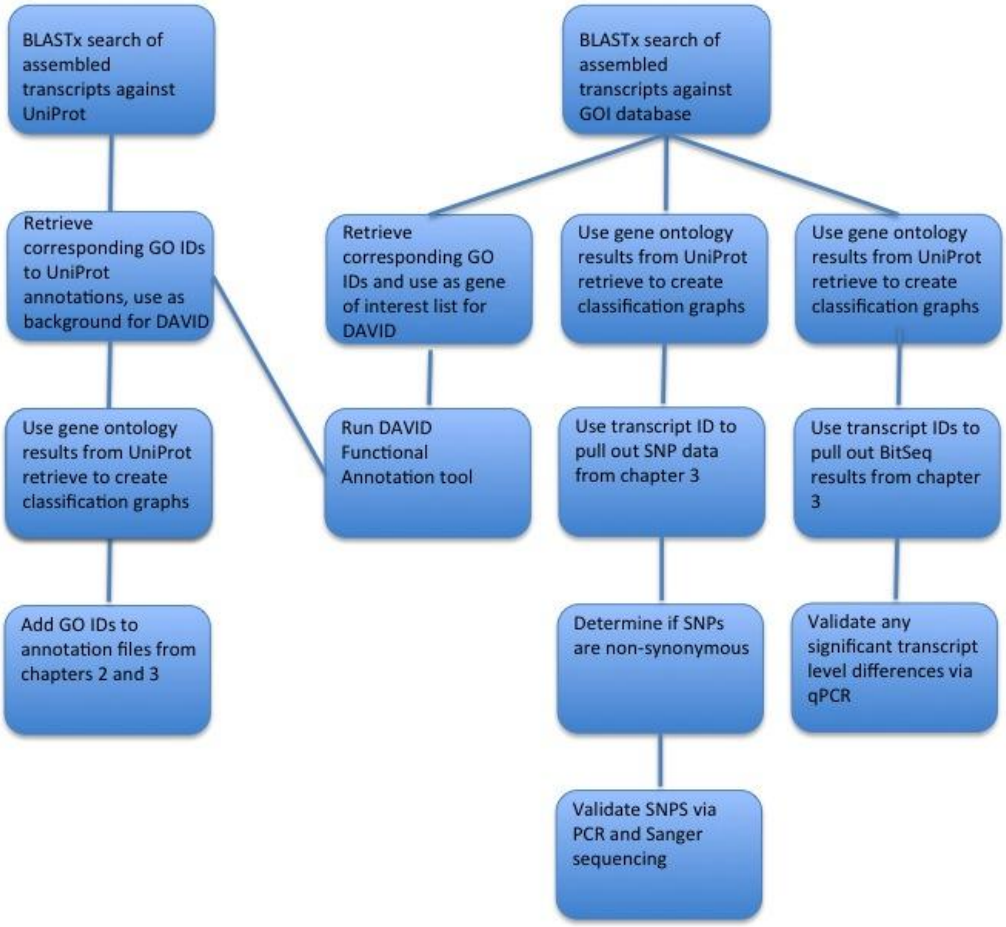
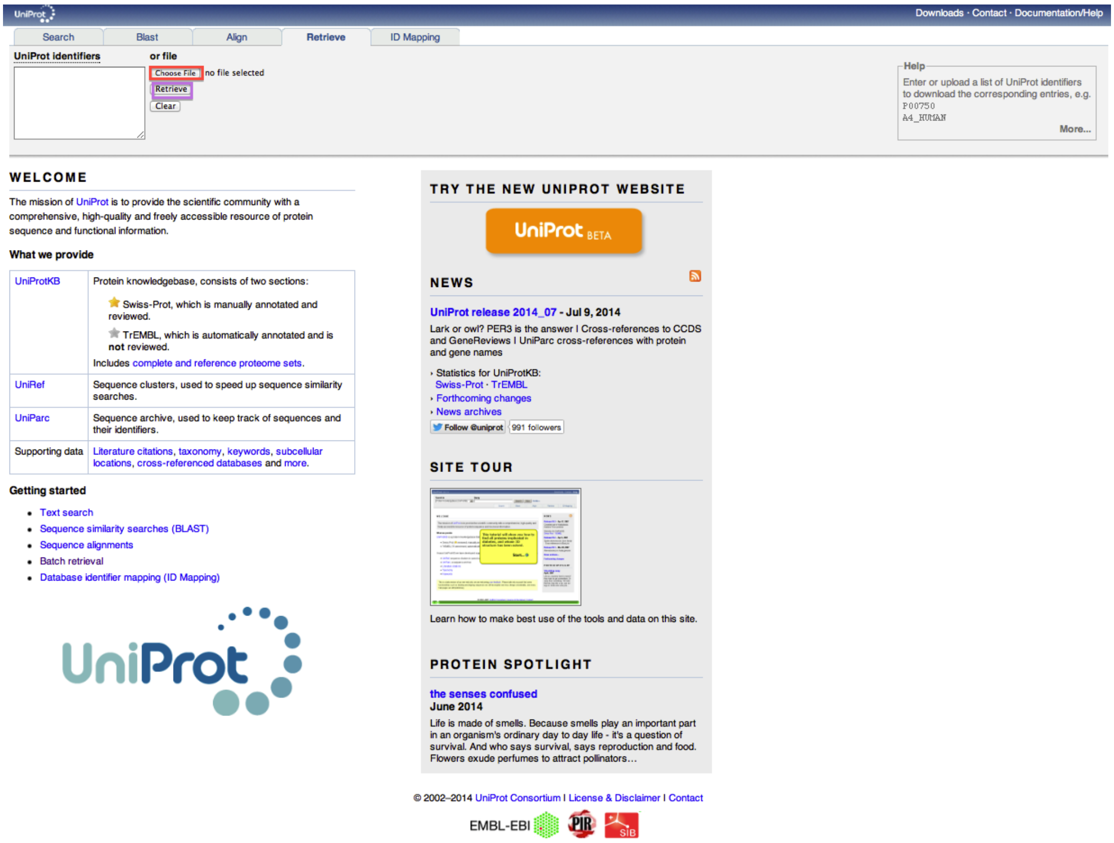

BLASTx search of assembled transcripts against UniProt
BLASTx search of assembled transcripts against GOI database
Retrieve corresponding GO IDs to UniProt annotations, use as background for DAVID
Retrieve corresponding GO IDs and use as gene of interest list for DAVID
Use gene ontology results from UniProt retrieve to create classification graphs
Use gene ontology results from UniProt retrieve to create classification graphs
Use gene ontology results from UniProt retrieve to create classification graphs
Run DAVID Functional Annotation tool
Use transcript ID to pull out SNP data from chapter 3
Use transcript IDs to pull out BitSeq results from chapter 3
Add GO IDs to annotation files from chapters 2 and 3
Determine if SNPs are non-synonymous
Validate any significant transcript level differences via qPCR
Validate SNPS via PCR and Sanger sequencing

Figure 4-9 Workflow of data analysis.

The steps are described in detail in sections 4.2.1 to 4.2.5.5 only the main steps are shown as a guide.

4.2.1 GO annotation

A conversion table of UniProt accessions and corresponding GO terms was produced using the UniProt retrieval program via the web interface following the steps laid out in figures 4.10 to 4.12. The UniProt accessions assigned to the three assemblies during the BLASTx annotation pipeline from chapters two and three were used as input.



UniProt identifiers

or file

no file selected

Help
Enter or upload a list of UniProt identifiers to download the corresponding entries, e.g.
P00750
A4_HUMAN
More...

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: ★ Swiss-Prot, which is manually annotated and reviewed. ☆ TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

TRY THE NEW UNIPROT WEBSITE

NEWS

UniProt release 2014_07 - Jul 9, 2014
Lark or owl? PER3 is the answer I Cross-references to CCDS and GeneReviews I UniParc cross-references with protein and gene names

› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives
Follow @uniprot 991 followers

SITE TOUR

Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

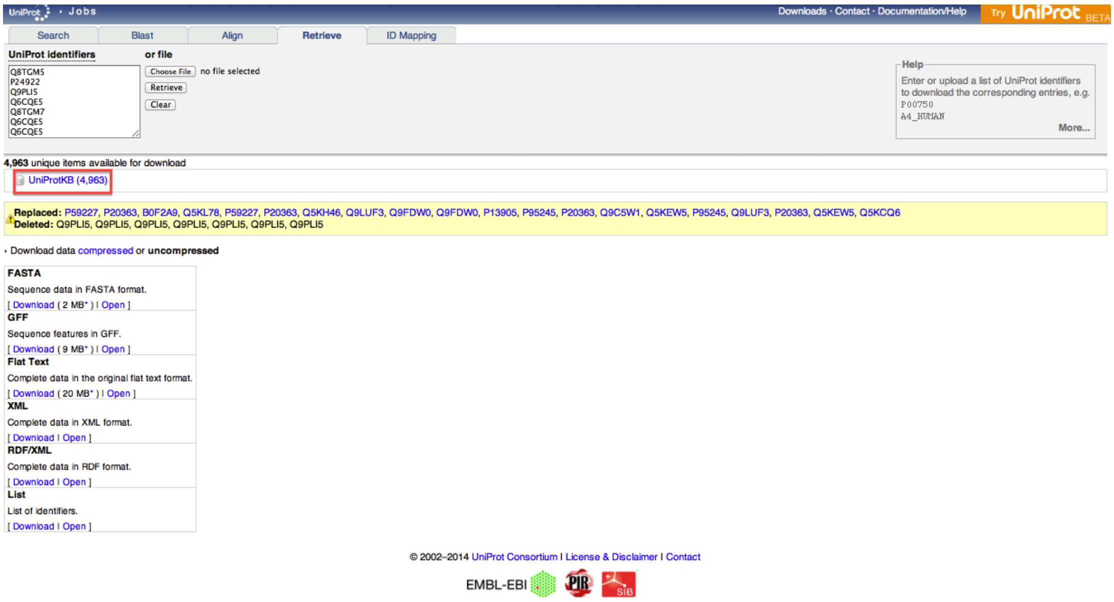
the senses confused
June 2014
Life is made of smells. Because smells play an important part in an organism's ordinary day to day life - it's a question of survival. And who says survival, says reproduction and food. Flowers exude perfumes to attract pollinators…

© 2002–2014 UniProt Consortium I License & Disclaimer I Contact

EMBL-EBI PIR SIB

Figure 4-10 Screen shot of UniProt web interface for the retrieval program.

The UniProt IDs assigned to the three assemblies were used as input as a text file using the choose file option (red box) and then the retrieve button (purple box) was used to pull out the corresponding entries from the database. (www.uniprot.org accessed in 2013).



UniProt identifiers
Q8TGM5
P24922
Q9PLI5
Q6CQE5
Q8TGM7
Q6CQE5
Q6CQE5

or file

no file selected

Help
Enter or upload a list of UniProt identifiers to download the corresponding entries, e.g.
P00750
A4_HUMAN
More...

4,963 unique items available for download

UniProtKB (4,963)

Replaced: P59227, P20363, B0F2A9, Q5KL78, P59227, P20363, Q5KH46, Q9LUF3, Q9FDW0, Q9FDW0, P13905, P95245, P20363, Q9C5W1, Q5KEW5, P95245, Q9LUF3, P20363, Q5KEW5, Q5KCQ6
Deleted: Q9PLI5, Q9PLI5, Q9PLI5, Q9PLI5, Q9PLI5, Q9PLI5, Q9PLI5

› Download data compressed or **uncompressed**

FASTA
Sequence data in FASTA format.
[Download (2 MB*) I Open]

GFF
Sequence features in GFF.
[Download (9 MB*) I Open]

Flat Text
Complete data in the original flat text format.
[Download (20 MB*) I Open]

XML
Complete data in XML format.
[Download I Open]

RDF/XML
Complete data in RDF format.
[Download I Open]

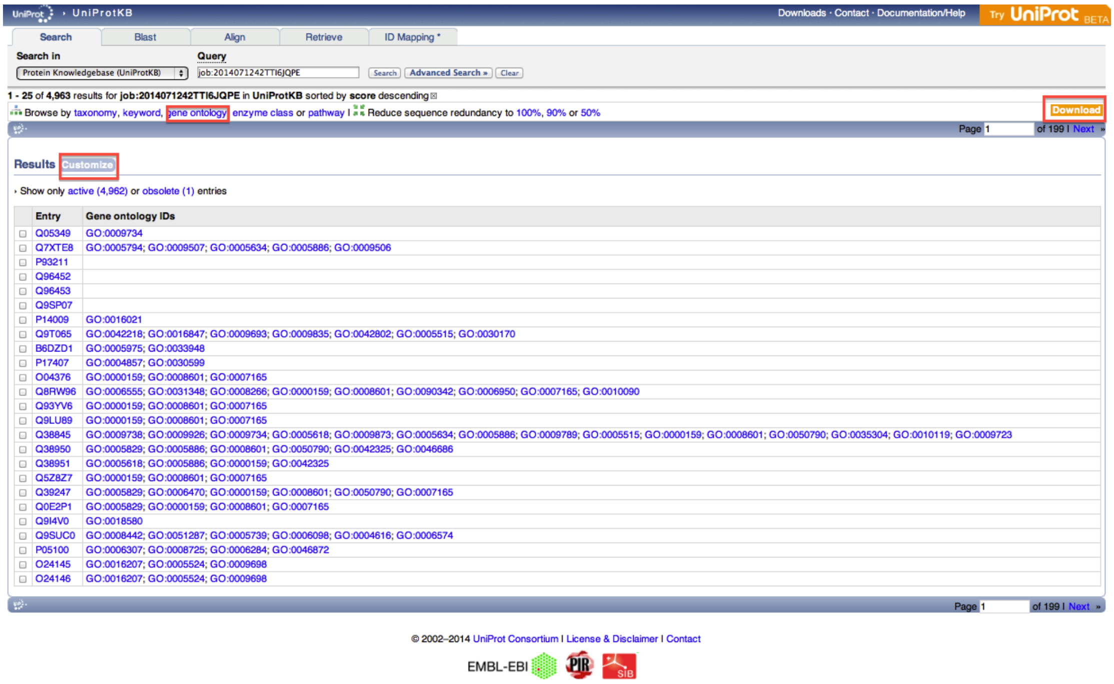
List
List of identifiers.
[Download I Open]

© 2002–2014 UniProt Consortium I License & Disclaimer I Contact

EMBL-EBI PIR SIB

Figure 4-11 The initial results page from the UniProt retrieval program.

The UniProtKB button (red box) is clicked to give the results table.



Search in: Protein Knowledgebase (UniProtKB)

Query: job:2014071242TTI6JQPE

1 - 25 of 4,963 results for job:2014071242TTI6JQPE in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway I Reduce sequence redundancy to 100%, 90% or 50%

› Show only active (4,962) or obsolete (1) entries

	Entry	Gene ontology IDs
□	Q05349	GO:0009734
□	Q7XTE8	GO:0005794; GO:0009507; GO:0005634; GO:0005886; GO:0009506
□	P93211	
□	Q96452	
□	Q96453	
□	Q9SP07	
□	P14009	GO:0016021
□	Q9T065	GO:0042218; GO:0016847; GO:0009693; GO:0009835; GO:0042802; GO:0005515; GO:0030170
□	B6DZD1	GO:0005975; GO:0033948
□	P17407	GO:0004857; GO:0030599
□	O04376	GO:0000159; GO:0008601; GO:0007165
□	Q8RW96	GO:0006555; GO:0031348; GO:0008266; GO:0000159; GO:0008601; GO:0090342; GO:0006950; GO:0007165; GO:0010090
□	Q93YV6	GO:0000159; GO:0008601; GO:0007165
□	Q9LU89	GO:0000159; GO:0008601; GO:0007165
□	Q38845	GO:0009738; GO:0009926; GO:0009734; GO:0005618; GO:0009873; GO:0005634; GO:0005886; GO:0009789; GO:0005515; GO:0000159; GO:0008601; GO:0050790; GO:0035304; GO:0010119; GO:0009723
□	Q38950	GO:0005829; GO:0005886; GO:0008601; GO:0050790; GO:0042325; GO:0046686
□	Q38951	GO:0005618; GO:0005886; GO:0000159; GO:0042325
□	Q5Z8Z7	GO:0000159; GO:0008601; GO:0007165
□	Q39247	GO:0005829; GO:0006470; GO:0000159; GO:0008601; GO:0050790; GO:0007165
□	Q0E2P1	GO:0005829; GO:0000159; GO:0008601; GO:0007165
□	Q9I4V0	GO:0018580
□	Q9SUC0	GO:0008442; GO:0051287; GO:0005739; GO:0006098; GO:0004616; GO:0006574
□	P05100	GO:0006307; GO:0008725; GO:0006284; GO:0046872
□	O24145	GO:0016207; GO:0005524; GO:0009698
□	O24146	GO:0016207; GO:0005524; GO:0009698

Figure 4-12 The final output needed to assign GO terms to the transcripts.

The table shown is the output for the 454 assembly for reference, the output can be customized to show different results such as IDs, GO terms, descriptions etc. via the customize button (red box). The table can be downloaded in several formats using the download button (red box).

The table was then used as a hash reference in a Perl script to add the GO terms to the annotation files from sections 2.4.3 and 3.4.2.15. To compare assemblies, GO terms were used to functionally categorize all transcripts on molecular function, cellular component and biological process by browsing the results by gene ontology.

4.2.2 Creation of the GOI (gene of interest) database from genes involved in secondary metabolite production in selected medicinally important plants to search for orthologs in daffodils

This was carried out following a scientific literature analysis (section 4.13 and 4.14). In addition to the plant systems discussed in section 4.14 it was decided to also examine *Arabidopsis thaliana,* as although it does not synthesize alkaloids, it is the model organism for plant study and was used for the annotation of the *Narcissus pseudonarcissus var.* Carlton reference. The literature search resulted in fourteen enzyme types linked to alkaloid biosynthesis for further investigation (see table 4.4). The FASTA sequences of these enzyme types from 18 plants were collected from the UniProt database (www.UniProt.org). A total of 159 fasta sequences were extracted from UniProt (see appendix section 6.6) and used to produce a BLAST database.

Table 4-4 The fourteen enzymes of interest used in the GOI database.

Pathways these enzymes have been associated with and plant sources.

Enzyme	Pathways	Plants with similar or exact enzyme	References
Tyrosine aminotransferase (TyrAT)	Ubiquinone and other terpenoid-quinone biosynthesis, cysteine and methionine metabolism, tyrosine metabolism, phenylalanine metabolism, phenylalanine, tyrosine and tryptophan biosynthesis and biosynthesis of amino acids.	*Papaver somniferum, Arabidopsis, Amaranthus caudatus, Chenopodium quinoa, Anchusa offinalis, Coleus blumei, Salvia miltiorrhiza and Capsicum annuum.*	114,253,259-261
Tyrosine decarboxylase (TDC)	Tyrosine metabolism, Isoquinoline alkaloid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Arabidopsis thaliana, Papaver somniferum.*	262-265
Norcoclaurine synthase (NCS)	Isoquinoline alkaloid biosynthesis, metabolic pathways, secondary metabolite biosynthesis.	*Arabidopsis thaliana, Papaver somniferum, Coptis japonica* and *Thalictrum flavum.*	18,203,266-270
6-O-methyltransferase (6OMT)	Isoquinoline alkaloid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Papaver somniferum, Thalictrum flavum, Coptis japonica.*	228,229,271
Coclaurine N-methyltransferase (CNMT)	Isoquinoline alkaloid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Papaver somniferum, Coptis japonica, Thalictrum flavum.*	121,272,273
(S)-N-methylcoclaurine-3'-hydroxylase (NMCH) (CYP80B)	Isoquinoline alkaloid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Papaver somniferum, Eschscholzia californica, Coptis chinensis, Papaver nudicaule, Papaver bracteatum,*	217,231,274-276
4'-O-methyltransferase (4OMT)	Isoquinoline alkaloid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Papaver somniferum, Eschscholzia californica, Coptis japonica.*	277,278
Hydroxycinnamoyl-CoA: Tyramine N-(hydroxycinnamoyl)transferase (THT)	Hydroxycinnamoyl-amine biosynthesis, phenylpropanoid biosynthesis, flavonoid biosynthesis and secondary metabolite biosynthesis.	*Papaver somniferum, Nicotiana tabacum, Solanum tuberosum, Capsicun annuum, Arabidopsis thaliana.*	279-281
Phenylalanine ammonia-lyase (PAL)	Phenylalanine metabolism, phenylpropanoid biosynthesis, metabolic pathways, secondary metabolite biosynthesis.	*Arabidopsis thaliana, Papaver somniferum and Capsicum annuum.*	247,260,282,283
4-hydroxyphenylpyruvate reductase (HPPR)	Ubiquinone and other terpenoid-quinone biosynthesis, tyrosine metabolism, phenylalanine metabolism and metabolic	*Papaver somniferum and Coleus blumei.*	114,284

	pathways.		
Cinnamic acid 4-hydroxylase (C4H)	Phenylalanine metabolism, phenylpropanoid biosynthesis, flavonoid biosynthesis, metabolic pathways and biosynthesis of secondary metabolites.	*Arabidopsis thaliana, Capsicum annuum* and *Salvia miltiorrhiza.*	20,254,285
4-coumarate CoA ligase (4CL)	Ubiquinone and other terpenoid-quinone biosynthesis, phenylalanine metabolism, phenylpropanoid biosynthesis, metabolic pathways and secondary metabolite biosynthesis.	*Arabidopsis thaliana and Capsicum annuum.*	247,286
Chalcone synthase (CHS)	Flavonoid and secondary metabolite biosynthesis.	*Arabidopsis thaliana, Capsicum annuum, Coleus blumei.*	287-289
Catechol-O-methyltransferase (COMT)	Steroid hormone biosynthesis, tyrosine metabolism, betalain biosynthesis, metabolic pathways.	*Capsicum annuum, Papaver somniferum, Arabidopsis thaliana* and numerous *Lamiaceae* herbs *(Ocimum basillicum, Salvia miltiorrhiza and Mentha piperita).*	228,265,290

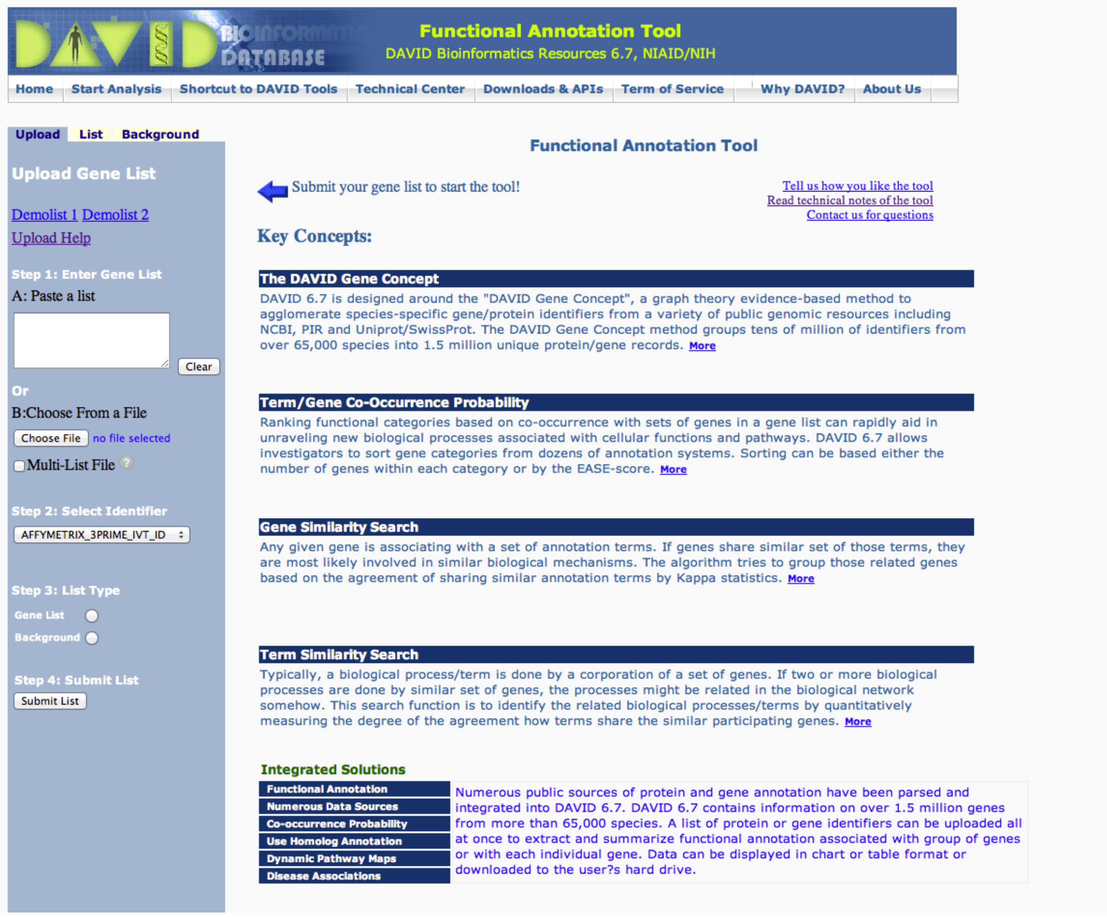
4.2.3 BLASTx search against GOI database for the 454 and post TE removal Trinity assemblies for putative transcripts involved in alkaloid production

The fasta files from all three assemblies were used in a BLASTX search against the GOI database and only the top hit for each transcript was carried through to the next step of the analysis. The BLAST search was carried out using the same settings as those used in chapter three, with the output set to tabulated and one hit per transcript.

4.2.4 GO enrichment studies for the GOI predicted transcripts

4.2.4.1 DAVID prediction of enrichment, EC and KEGG annotation

DAVID v6.7 was used with the 454 and both Trinity (Original and Post-TE removal) assemblies to look for enrichment within the genes predicted to be daffodil orthologs of those in the GOI database (section 4.2.2). The functional annotation tool within DAVID was used via its web interface (figure 4.13). This tool requires UniProt accessions and the BLAST results from the UniProt annotation (2.3.4.5 and 3.4.2.15) were used as background lists for both assemblies (figure 4.13). The transcripts that matched those in the GOI database were used as the gene list for the analysis. Within the annotation summary results the default settings were used, with addition of the EC number.



Functional Annotation Tool

Submit your gene list to start the tool!

Tell us how you like the tool
Read technical notes of the tool
Contact us for questions

Key Concepts:

The DAVID Gene Concept

DAVID 6.7 is designed around the "DAVID Gene Concept", a graph theory evidence-based method to agglomerate species-specific gene/protein identifiers from a variety of public genomic resources including NCBI, PIR and Uniprot/SwissProt. The DAVID Gene Concept method groups tens of million of identifiers from over 65,000 species into 1.5 million unique protein/gene records. More

Term/Gene Co-Occurrence Probability

Ranking functional categories based on co-occurrence with sets of genes in a gene list can rapidly aid in unraveling new biological processes associated with cellular functions and pathways. DAVID 6.7 allows investigators to sort gene categories from dozens of annotation systems. Sorting can be based either the number of genes within each category or by the EASE-score. More

Gene Similarity Search

Any given gene is associating with a set of annotation terms. If genes share similar set of those terms, they are most likely involved in similar biological mechanisms. The algorithm tries to group those related genes based on the agreement of sharing similar annotation terms by Kappa statistics. More

Term Similarity Search

Typically, a biological process/term is done by a corporation of a set of genes. If two or more biological processes are done by similar set of genes, the processes might be related in the biological network somehow. This search function is to identify the related biological processes/terms by quantitatively measuring the degree of the agreement how terms share the similar participating genes. More

Integrated Solutions

- Functional Annotation
- Numerous Data Sources
- Co-occurrence Probability
- Use Homolog Annotation
- Dynamic Pathway Maps
- Disease Associations

Numerous public sources of protein and gene annotation have been parsed and integrated into DAVID 6.7. DAVID 6.7 contains information on over 1.5 million genes from more than 65,000 species. A list of protein or gene identifiers can be uploaded all at once to extract and summarize functional annotation associated with group of genes or with each individual gene. Data can be displayed in chart or table format or downloaded to the user?s hard drive.

Figure 4-13 A screen shot of the DAVID web interface for the functional annotation tool.

The background is set as UniProt accessions from the UniProt BLAST results from the annotation step and the gene list is UniProt accessions from the predicted gene BLAST results. (http://david.abcc.ncifcrf.gov/summary.jsp)

From this stage onwards, only results from mapping the post TE removal Illumina reads to both the 454 and Post-TE removal assemblies were used for further analysis.

4.2.5 The determination of non-synonymous SNPS within the putative transcripts- Prediction of ORF and non-synonymous SNPs

To identify SNPs within the predicted transcripts, the whole transcriptome SNP profiling results from chapter three (section 3.4.2.13) were cross-referenced with the contigs that resulted in a BLAST hit against the GOI database. Using the BLASTX alignment output format (-m 0) against the GOI database it is possible to predict the reading frame for the section of the transcript that hits the protein sequence of the gene of interest. Such polymorphisms could then be examined to determine whether the SNP was non-synonymous or synonymous. Any contigs predicted to have non-synonymous SNPs were then taken forward for Sanger sequencing validation (section 4.2.6.4).

4.2.6 The determination of putative transcripts with significant differences in transcript levels between daffodil varieties

4.2.6.1 BitSeq

The whole transcriptome transcript level profiling results from chapter three were cross-referenced with the transcript IDs that produced a hit against the GOI database. Any transcripts that showed significant transcript level differences (PPLR of ≤0.05 ≥ 0.95) were then taken forward to the validation stage of Sanger sequencing and qPCR (sections 4.2.6.4 and 4.2.6.5).

4.2.6.2 Confirmation of transcripts via RT-PCR and transcript level differences via RT-qPCR

Template cDNA was prepared from three different basal plates for each variety; the RNA was extracted as described in section 3.3.1.1. The RNA was converted to cDNA using the Qiagen Quantitech Reverse Transcription Kit following the manufacturer's instructions and then diluted to 100µl (x5 dilution). Transcript specific primers were designed according to the Carlton assembled reads using Primer3. (http://primer3.ut.ee) (See section 6.7 of appendix).

4.2.6.3 Confirmation of predicted transcripts

PCR was carried out on one biological sample (3 technical replicates) per variety using Bioline MyTaq ™ Red Mix in 10µl reactions according to the manufacturer's protocol. The PCR program was as follows: 94°C for 1 minute followed by 35 cycles of amplification (20s denaturing at 94°C, 20s at annealing temperature, and 30s extension at 72°C) and a final extension for 5 minutes at 72°C.

Initial results showed two or more bands for some transcripts and so nested primers were designed (see section 6.7 of appendix) and used under the same PCR conditions as above.

Some primer pairs (see section 6.7 of appendix) had dissimilar annealing temperatures and so touchdown PCR was carried out as follows: 95°C for 3 minutes, 95°C for 30s followed by 9 cycles starting at 70°C for 45s, annealing temperature for 60s then 70°C for 45s. In the first cycle the annealing temperature was 70°C, and in each following cycle the annealing temperature was lowered by 2°C until a final cycle at 52°C. After this there were 25 cycles of amplification (95°C for 30s, 50°C for 45s and 72°C for 1 minute) followed by 72°C for 5 minutes. The PCR products were separated on 1% agarose gels to confirm size and those that were predicted to have non-synonymous SNPs were sent for Sanger sequencing confirmation.

4.2.6.4 Sanger sequencing to confirm SNPs

DNA amplified from selected transcripts with predicted non-synonymous SNPs was sent for Sanger sequencing. Each PCR product was used for one sequencing reaction using the forward primer and one using the reverse. The PCR product clean up and sequencing was carried out by the Source Bioscience Life Sciences Sanger Sequencing services.

(http://www.lifesciences.sourcebioscience.com/genomic-services/sanger-sequencing-service/).

4.2.6.5 Real time qPCR of those transcripts predicted to have significant transcript level differences

In order to quantify the transcript level differences further, RT-qPCR was carried out, using actin as a control as it showed no significant difference in BitSeq analysis between varieties. The relative actin response for each variety was used to standardize the transcript level differences of the transcripts of interest. For each transcript, three biological and three technical replicates were measured for each replication a negative control from the cDNA preparation without the reverse transcriptase was used as well as using a control with nuclease free water instead of the template cDNA. These controls were used to check for genomic DNA contamination. (See section 6.7 in appendix for plate layout).

The RT-qPCR was carried out using the Bioline SensiFast™ SYBR® Hi-ROX Kit and an Agilent Technologies Stratagene MX3005P PCR machine. The conditions were as follows: 95°C for 2 minutes followed by 40 cycles of 95°C for 5 seconds, annealing temperature for 22 seconds and 72°C for 15 seconds.

To work out a fold change (relative to actin) the following is determined for each biological rep:

$$x = 2^{(S-A)}$$

Where S= average sample CT (threshold cycle) from three technical reps and A= average actin CT from three technical reps.

The average of the three biological reps is then worked out and used to calculate fold change:

$$fold\ change = \frac{(\frac{x_{AC1} + x_{AC2} + x_{AC3}}{3})}{(\frac{x_{C1} + x_{C2} + x_{C3}}{3})}$$

Where AC1, AC2 and AC3 are the three biological replicates for Andrew's choice and C1, C2 and C3 are the three biological replicates for Carlton.

4.3 Results

4.3.1 Whole transcriptome functional annotation

4.3.1.1 Functional categorization of the 454 assembly

The GO annotation via UniProt resulted in 4147 transcripts being assigned to molecular function, with the top three categories being binding (73%), catalytic activity (62%) and transporter activity (8%). A total of 4118 were assigned a cellular component category with the top three being cell (92%), cell part (92%) and organelle (72%). Finally, 4421 were assigned a biological process category and the top three categories were metabolic processes (81%), cellular processes (80%) and single-organism processes (61%). The percentage given is the percentage of the total assigned IDs linked to this category. An overview of the assignments of the three annotation modules can be seen in figures 4.14, 4.17 and 4.20.

4.3.1.2 Functional categorization of the Original Trinity assembly

The GO annotation via UniProt resulted in 9965 transcripts being assigned to molecular function, with the top three categories being binding (74%), catalytic activity (63%) and transporter activity (7%). A total of 10066 were assigned a cellular component category with the top three being cell or cell part (91%), organelle (71%) and organelle part (39%). Finally, 10797 were assigned a biological process category and the top three categories were cellular processes (81%), metabolic processes (80%) and single-organism processes (61%). The breakdown of the three annotation modules can be seen in figures 4.15, 4.18 and 4.21.

4.3.1.3 Functional categorization of the Post-TE removal Trinity assembly

The GO annotation via UniProt resulted in 10023 transcripts being assigned to molecular function, with the top three categories being binding (74%), catalytic activity (63%) and transporter activity (7%). A total of 10098 were assigned to a cellular component category with the top three being cell or cell part (91%), organelle (71%) and organelle part (39%). Finally 10845 were assigned a biological process category, the top three categories were cellular processes (81%), metabolic processes (80%) and single-organism processes (61%). The breakdown of the three annotation modules can be seen in figures 4.16, 4.19 and 4.22.

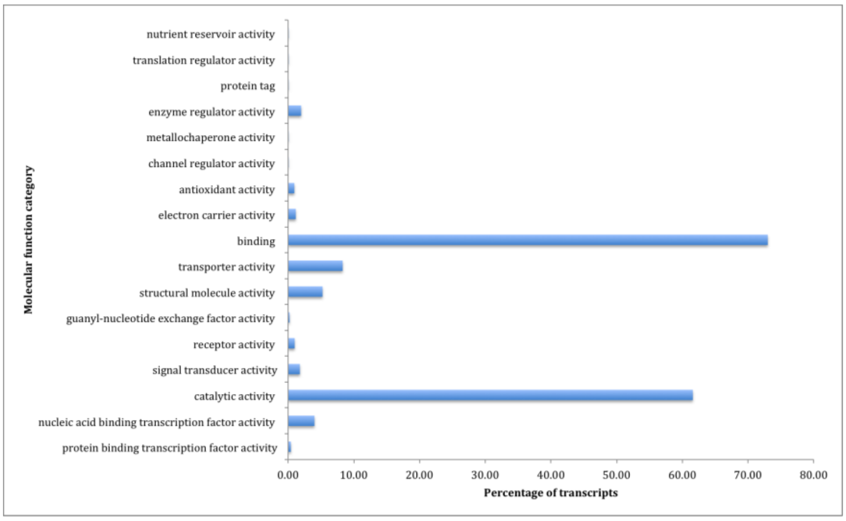


Figure 4-14 Molecular function assignment for the UniProt GO analysis of the 454 assembly.

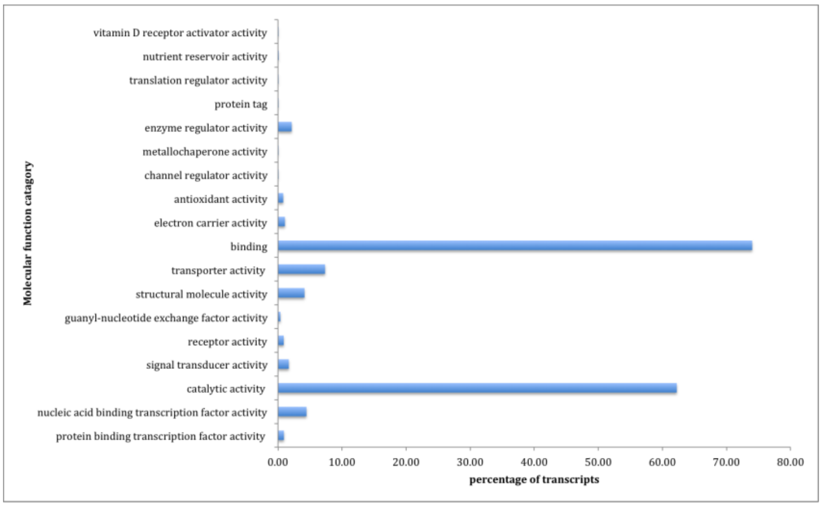


Figure 4-15 Molecular function assignment for the UniProt GO analysis for the Original Trinity assembly.

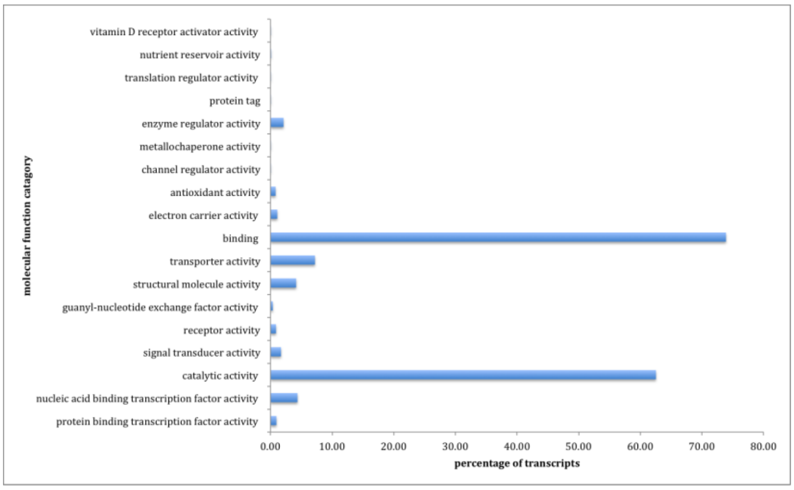


Figure 4-16 Molecular function assignment for the UniProt GO analysis of the Post-TE removal Trinity assembly.

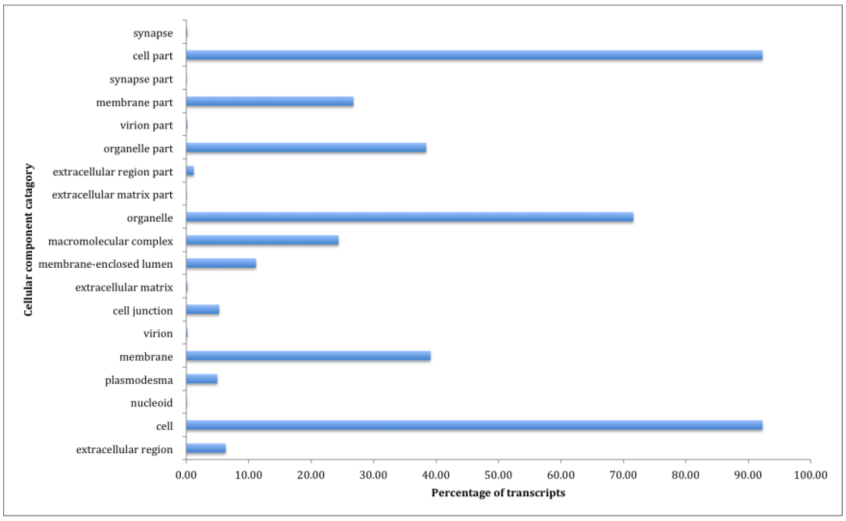


Figure 4-17 Cellular Component assignment from the UniProt GO categorization for the 454 assembly.

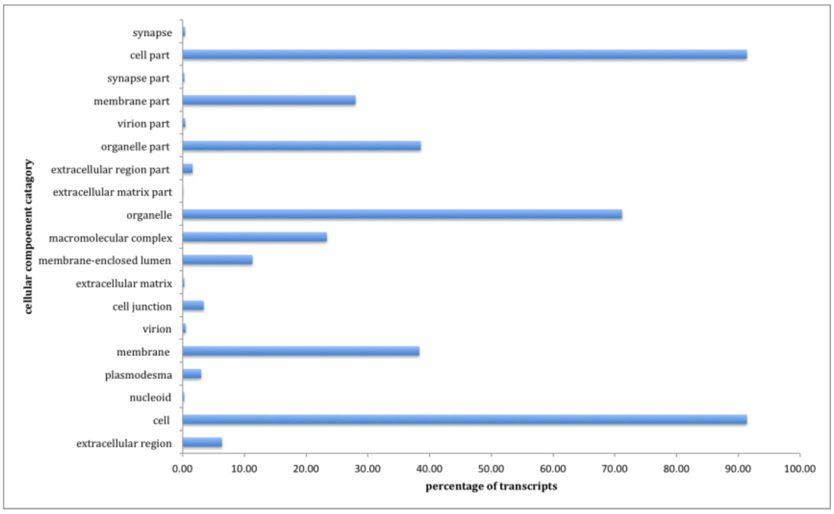


Figure 4-18 Cellular component assignment from the UniProt GO analysis for the Original Trinity assembly.

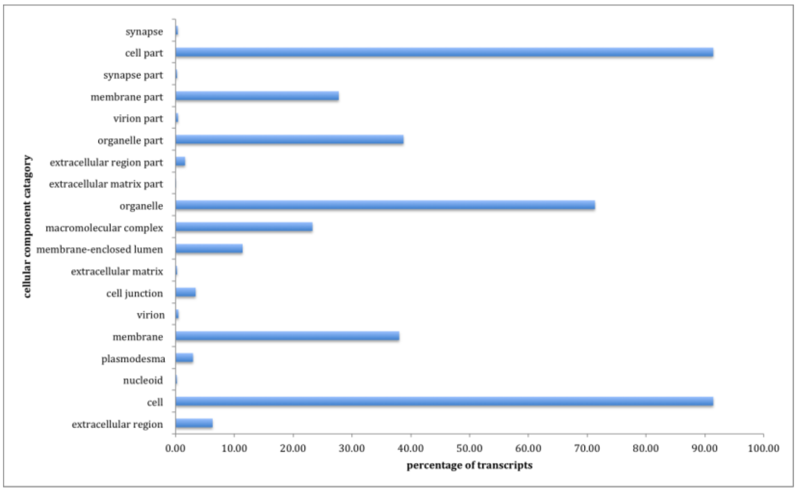


Figure 4-19 Cellular Component assignment for the UniProt GO analysis of the Post-TE removal Trinity Assembly.

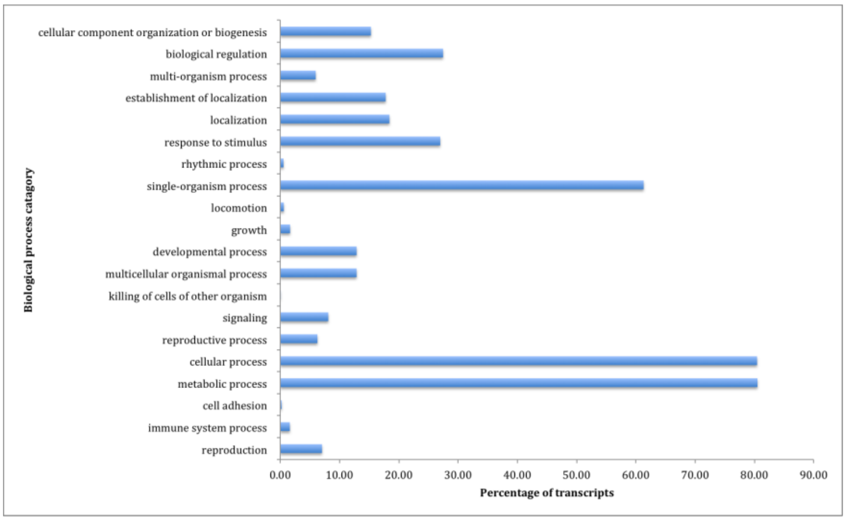


Figure 4-20 Biological Process assignment from the UniProt GO analysis for the 454 assembly.

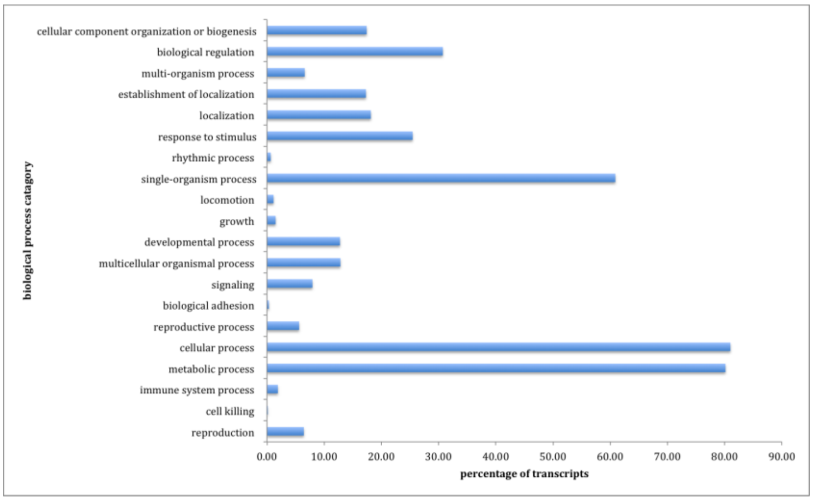


Figure 4-21 Biological Process assignment for the UniProt GO analysis for the Original Trinity assembly.

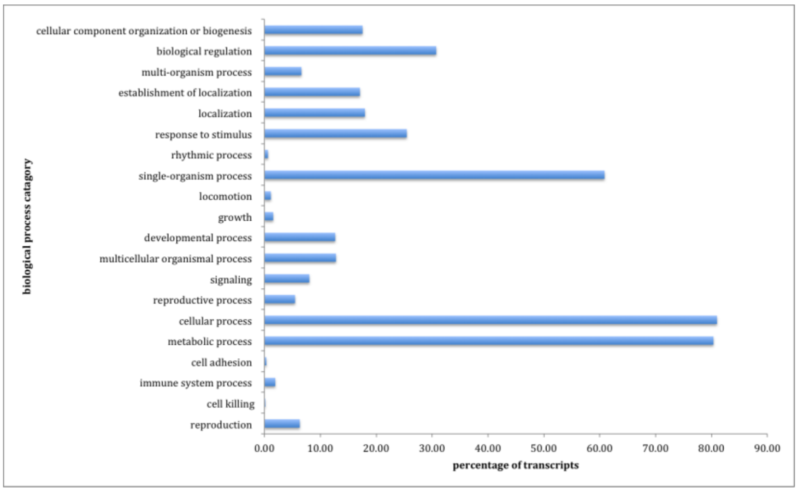


Figure 4-22 Biological Process assignment for the UniProt GO analysis of the Post-TE removal assembly.

4.3.1.4 The assignment of GO terms to transcripts

The UniProt GO annotation resulted in 5383 transcripts (68% of the 7729 transcripts that were annotated with UniProt IDs (section 4.2.1)) being assigned GO IDs. A larger number of transcripts (12377) from the Original Trinity assembly with UniProt IDs were assigned GO annotation. However, they comprised a smaller proportion of the total number of transcripts in this assembly (29% of the 42335 assigned UniProt IDs) than for the 454 assembly. A slightly higher number (15808) and proportion (38%) of the Post-TE removal Trinity assembly were assigned GO terms. Of these 9556 matched transcripts from the Original assembly and 6252 were new transcripts annotated after the TE removal. The first five entries, arranged in random order, in each of these annotation lists are shown in tables 4.5-4.7 the whole data sets can be viewed on the appendix disc.

Table 4-5 GO annotation results for the 454 assembly.

Uniprot ID	GO ID	Daffodil transcript ID
Q5PQL3	GO:0030660; GO:0042500; GO:0010008; GO:0071458; GO:0071556; GO:0005765; GO:0006509; GO:0031293; GO:0005886; GO:0042803; GO:0050776	HDA57HA01ANGJN HDA57HA01ARE2O
Q7Q161	GO:0050662; GO:0006098; GO:0004616	Contig03905
Q9JKB1	GO:0007628; GO:0032869; GO:0005737; GO:0042755; GO:0008233; GO:0045600; GO:0030163; GO:0016579; GO:0060041; GO:0006511; GO:0004843	HDA57HA01A0EC7
Q6DEU9	GO:0016593; GO:0001711; GO:0080182; GO:0045638; GO:0000122; GO:0051571; GO: 2001162; GO:0032968; GO:0045944; GO:0019827; GO:0006351; GO:0035327	HDA57HA01A2Q1V
O64629	GO:0005524; GO:0000775; GO:0043987; GO:0043988; GO:0035175; GO:0044022; GO:0016572; GO:0005634; GO:0048471; GO:0005819	HDA57HA01BLCDO

Table 4-6 The GO annotation results for the Original Trinity assembly.

Uniprot Id	GO ID	Daffodil Transcript ID
Q9LZB8	GO:0005524; GO:0006200; GO:0042626; GO:0010541; GO:0010315; GO:0010329; GO:0010540; GO:0009507; GO:0009941; GO:0031969; GO:0016021; GO:0005886; GO:0055085; GO:0005215	Daff105662 Daff107641 Daff46405 Daff46406
Q9JKB1	GO:0007628; GO:0032869; GO:0005737; GO:0042755; GO:0008233; GO:0045600; GO:0030163; GO:0016579; GO:0060041; GO:0006511; GO:0004843	Daff106157
Q9CK84	GO:0006354; GO:0006353; GO:0032784; GO:0031564	Daff139665
O64629	GO:0005524; GO:0000775; GO:0043987; GO:0043988; GO:0035175; GO:0044022; GO:0016572; GO:0005634; GO:0048471; GO:0005819	Daff109922 Daff131772
Q94BM7	GO:0005524; GO:0080008; GO:0042802; GO:0005634; GO:0005515; GO:0004672; GO:0009585	Daff113015 Daff117266 Daff23517 Daff48165 Daff48166

Table 4-7 The GO annotation for the Post-TE removal Trinity assembly.

Uniprot ID	GO ID	Daffodil transcript ID
Q5PQL3	GO:0030660; GO:0042500; GO:0010008; GO:0071458; GO:0071556; GO:0005765; GO:0006509; GO:0031293; GO:0005886; GO:0042803; GO:0050776	Daff93414
Q9LZB8	GO:0005524; GO:0006200; GO:0042626; GO:0010541; GO:0010315; GO:0010329; GO:0010540; GO:0009507; GO:0009941; GO:0031969; GO:0016021; GO:0005886; GO:0055085; GO:0005215	Daff105662 Daff107641 Daff46405 Daff46406
Q9JKB1	GO:0007628; GO:0032869; GO:0005737; GO:0042755; GO:0008233; GO:0045600; GO:0030163; GO:0016579; GO:0060041; GO:0006511; GO:0004843	Daff106157
Q9CK84	GO:0006354; GO:0006353; GO:0032784; GO:0031564	Daff139665
Q1GFB5	GO:0005524; GO:0050567; GO:0006412	Daff58758

4.3.2 BLASTx search against the GOI database for the 454 and pre/Post-TE removal Trinity assembled transcripts for putative transcripts involved in alkaloid production

The total number of transcripts predicted to be orthologs of known secondary metabolite biosynthesis genes is 111 for the 454 data and 461 for the Original Trinity assembly. A similar number (448) were predicted as orthologs in the Post-TE removal Trinity assembly of which 281 matched to the pre–TE removal assembly and 166 were previously unseen transcripts.

4.3.3 Pathway analysis: the use of DAVID to predict enrichment from EC and KEGG annotation

4.3.3.1 DAVID analysis for gene enrichment

All three assemblies resulted in very similar DAVID annotation clustering and an overview showing the enrichment scores and most common ID terms are given in tables 4.8, 4.9 and 4.10. The 454 assembly had considerably fewer clusters (14 compared to 27 in the Trinity assemblies) but this is to be expected with its lower number of transcripts (77 submitted as the GOI list compared to 200 and 201 for the pre- and Post-TE Trinity assemblies). The clustering tables for the full functional annotation for all three assemblies can be seen on the appendix disc. All three assemblies show the highest enrichment score for a cluster of CYP450s, specifically those linked to oxidoreductase. In all three assemblies this term has the highest percentage of genes in the gene of interest list, ~50% for each assembly (454 54%, Original 51% and Post-TE 48.7%) and a fold enrichment score of ~8 for each assembly (454 7.4, Original 8 and post TE 8.5).

The other clusters are also similar for all three assemblies. Interesting, a cluster linked to PAL (an enzyme known to be involved in galanthamine production) is present in all three with a significant enrichment score (454 2.71, Original 6.08, Post-TE 6.1). The term “phenylalanine ammonia-lyase” is linked to 6% of the total number of genes but has a FE of 65 for the 454 assembly, where as for both the Original and Post-TE removal assemblies it contains 3.8% of the total number of genes with a FE of 67 for the pre- and 71 for the Post-TE assembly.

Finally, in all three assemblies the term “secondary metabolism” is associated with both a high percentage of total genes and a significant fold enrichment score (454 23 % and 9.58 FE, Original 25% and 12 FE and Post-TE 25% and 12 FE).

Support for these assignments must also come from the EC, KEGG pathways and the relative representation of the genes showing high scores. These will be brought together in the Discussion (section 4.4.3).

Table 4-08 Summary of DAVID analysis for 454 assembly.

The enrichment score of each cluster is shown along with the key GO terms. An enrichment score is considered above 1.3. The full functional clustering can be seen on appendix disc.

Cluster Number	Enrichment score	Key terms
1	14.74	Cytochrome P450, oxidation reduction
2	8.79	Oxidoreductase
3	5.29	Methyltransferase
4	5.04	Secondary metabolism process, phenylpropanoid metabolism and biosynthesis, aromatic compound biosynthesis
5	4.22	Aminotransferase
6	2.91	Alkene biosynthesis
7	2.72	Phenylpropanoid metabolism
8	2.28	Lipid biosynthesis, organic/carbonic/fatty acid synthesis
9	1.86	Isopropenoid, terpenoid, diterpenoid metabolism
10	1.65	Ligase, ATP binding, nucleotide binding
11	1.58	Co-enzyme binding, nucleotide binding
12	1.3	Reproductive development
13	0.41	Response to stimuli
14	0.11	Chloroplast

Table 4-09 Summary of DAVID results for the Original Trinity assembly.

The enrichment score of each cluster is shown along with the key GO terms. An enrichment score is considered above 1.3. The full functional clustering can be seen on the appendix disc.

Cluster Number	Enrichment score	Key terms
1	38.82	Cytochrome P450, oxidoreductase, ion-binding

2	28.8	Secondary metabolism, phenylpropanoid metabolism and biosynthesis, amino acid biosynthesis
3	15.1	Methyltransferase
4	14.89	Oxidoreductase
5	14.89	Oxygen binding
6	11.03	Vitamin binding, aminotransferase
7	10.94	Fatty/organic/carboxylic/lipid acid metabolism
8	7.75	Methyltransferase
9	6.72	Fatty/lipid acid ligase
10	6.1	Phenylpropanoid metabolism
11	5.44	Co-A ligase, lipid/secondary metabolite metabolism
12	5.06	Ethylene biosynthesis/metabolism, reproductive development
13	3.84	Iso/di/terpenoid metabolism/biosynthesis
14	3.19	Glycosylate reductase activity, NAD/NADH binding, co-enzyme binding
15	2.89	Trans-cinnamate-4-monooxygenase activity
16	2.44	Microbody/peroxisome
17	2.24	Acyltransferase
18	1.85	Hormone biosynthesis/metabolism
19	1.79	Carboxylase activity
20	1.58	Vesicular, membrane, cell fraction
21	1.02	Homeostatic process
22	0.73	Fatty acid metabolism
23	0.51	Response to stimuli
24	0.43	Development, growth
25	0.09	Amino acid biosynthesis, chloroplast
26	0.02	Response to stimuli
27	0.002	Lumen

Table 4-10 Summary of DAVID analysis for the Post-TE Trinity assembly.

Cluster Number	Enrichment score	Key terms
1	35.13	Cytochrome P450, oxidoreductase, ion-binding
2	24.57	Secondary metabolism, phenylpropanoid metabolism and biosynthesis, amino acid biosynthesis
3	15.27	Methyltransferase
4	14.85	Oxidoreductase
5	10.25	Fatty/organic/carboxylic/lipid acid metabolism
6	9.57	Oxygen binding
7	8.04	Vitamin binding, aminotransferase, co-factor binding
8	7.7	Ligase activity, co-enzyme activity
9	7.56	Lignin biosynthesis/metabolism, O-methyltransferase
10	6.08	L-phenylalanine metabolism, aromatic compound catabolic process, aromatic amino acid family metabolism

11	5.07	Alkene biosynthesis, reproductive process
12	4.74	Co-A ligase activity, ATP/nucleotide binding
13	3.91	Iso/di/terpenoid metabolism/biosynthesis
14	3.13	Glycosylate reductase activity, NAD/NADH binding, co-enzyme binding
15	2.94	Trans-cinnamate-4-monooxygenase activity
16	2.51	Microbody/peroxisome
17	2.28	Acyltransferase
18	1.71	Carboxylase activity
19	1.14	Microsome, vesicular, cell fraction
20	1.06	Hormone biosynthesis
21	1.04	Homeostatic process
22	0.72	Mitochondrion
23	0.55	Response to stimuli
24	0.42	Development, growth
25	0.1	Amino acid biosynthesis, chloroplast
26	0.02	Response to stimuli
27	0.002	Lumen

The enrichment score of each cluster is shown along with the key GO. An enrichment score is considered above 1.3. The full functional clustering results are on appendix disc.

4.3.3.2 KEGG pathway annotation of GOI daffodil transcripts

Since the KEGG pathway data is used within the DAVID analysis, the annotation terms will mirror those in the enrichment score annotation. These are summarized in tables 4.14-4.16 below. Similar pathways from the GOI list will be represented in all three assemblies. The 454 assembly contained transcripts linked to phenylalanine and tyrosine metabolism (both 2) and also to the biosynthesis of plant hormones (4). The Original assembly predicted one transcript associated with the pathway "isoquinoline alkaloid biosynthesis" (DAVID ID 3291284, UniProt ID Q7XHL3, tyrosine decarboxylase 1). It also showed the highest number of transcripts linked to plant hormone biosynthesis (9) and also the two similar terms "biosynthesis of phenylpropanoids" (6) and "phenylpropanoid biosynthesis" (4). These differ only in that the biosynthesis of phenylpropanoids includes the DAVID IDs 3497234 (UniProt ID Q96330, flavonol synthase from *Arabidopsis*) and 2941875 (UniProt ID P51090 *Vitis vinifera* Chalcone synthase) in addition to the four located within phenylpropanoid biosynthesis. This will be discussed further in section 4.4.3. An unexpected result is the association of a transcript to pathways linked to cancer, "pathways in cancer" and "chronic myeloid leukemia" (DAVID ID 588768, UniProt ID Q0VCQ1, C-terminal binding protein in *Bos taurus*) which will also be discussed in section 4.4.3.

The Post-TE assembly showed very similar results to the Original assembly.

Table 4-8 Summary of KEGG results for the 454 data.

Pathways from different species with the same pathway name were grouped together.

KEGG pathway	UniProt IDs	Number of unique IDs assigned to it
Alpha-Linolenic acid metabolism	Q10S72	1
Fatty acid metabolism	Q9T0A0	1
Alanine, aspartate and glutamate metabolism	Q56232	1
Steroid biosynthesis	Q6ZIX2	1
Limonene and pinene degradation	Q9FG65	1
Biosynthesis of terpenoids and steroids	Q6ZIX2 Q42569 Q9C5Y2	3
Glycine, serine and threonine metabolism	O04130	1
Diterpenoid biosynthesis	Q9C5Y2 Q9XFR9	2
Glyoxylate and dicarboxylate metabolism	Q9SXP2	1
Lysine biosynthesis	Q93ZN9	1
Flavone and flavonol biosynthesis	Q9FK25	1
Ubiquinone and other terpenoid-quinone biosynthesis	Q9ZSK1 P04694	2
Arachidonic acid metabolism	Q2KIG5	1
Biosynthesis of plant hormones	Q10S72 Q6ZIX2 Q42569 Q9C5Y2	4
Phenylalanine metabolism	Q56232 P04694	2
Cysteine and methionine metabolism	Q56232 P04694	2
Stilbenoid, diarylheptanoid and gingerol biosynthesis	Q9FG65	1
Phenylalanine, tyrosine and tryptophan biosynthesis	Q56232 P04694	2
Methane metabolism	Q9SXP2	1
Novobiocin biosynthesis	Q56232	1
Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid	Q93ZN9	1
Arginine and proline metabolism	Q56232	1
Endocytosis	Q93ZN9	1
Tyrosine metabolism	Q56232 P04694	2

Table 4-9 KEGG summary for the Original Trinity assembly.

Those pathways with the same name but from different species were grouped together.

KEGG pathway	IDs	Number of unique IDs assigned to it
Biosynthesis of alkaloids derived from shikimate pathway	3510236 3491601 3264871 3291284	4
Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid	3512745 3491601	2
Biosynthesis of phenylpropanoids	3510236 3511325 3497234 3491601 3264871 2941875	6
Flavone and flavonol biosynthesis	3284426	1
Histidine metabolism	3291284	1
Tyrosine metabolism	4000758 3291284 384445	3
Phenylalanine metabolism	4000758 3491601 3291284 384445	4
Tryptophan metabolism	3516231 3291284	2
Isoquinoline alkaloid biosynthesis	3291284	1
Fatty acid metabolism	3491824 817215 3503787 3489853 3493055	5
PPAR signalling pathway	817215	1
Adipocytokine signalling pathway	817215	1
Glycolysis / Gluconeogenesis	5636578 825753	2
Pyruvate metabolism	5636578 825753	2
Propanoate metabolism	5636578 825753	2
Flavonoid biosynthesis	3497234 2941875	2
Circadian rhythm	2941875	1
Arachidonic acid metabolism	615542 597566	2
Ubiquinone and other terpenoid-quinone biosynthesis	3510236 3264871 384445	3
Cysteine and methionine metabolism	3274554 4000758 384445	3
Phenylalanine, tyrosine and tryptophan biosynthesis	4000758 384445	2
Biosynthesis of plant hormones	3274554 3266521 3490600 3277060 3511307 3286309 3499037 3509334 3491601	9

Phenylpropanoid biosynthesis	3510236 3511325 3491601 3264871	4
Alpha-Linolenic acid metabolism	3266521 3490600	2
Brassinosteroid biosynthesis	3499627 3271088 3501700 3286309	4
Limonene and pinene degradation	3507363 3496243 3500475 3506487 3503663 3502969 3494348 3497859	8
Stilbenoid, diarylheptanoid and gingerol biosynthesis	3507363 3496243 3500475 3506487 3503663 3502969 3494348 3497859	8
Carotenoid biosynthesis	3284113 3286008 3493449	3
Alanine, aspartate and glutamate metabolism	5636578 4000758	2
Arginine and proline metabolism	4000758	1
Glycine, serine and threonine metabolism	3948583 3498205	2
Diterpenoid biosynthesis	3499037 3509334 3493541 3517679	4
Biosynthesis of terpenoids and steroids	3511307 3499037 3509334	3
Glyoxylate and dicarboxylate metabolism	3218943 3929207	2
Lysine biosynthesis	3512745	1
Endocytosis	3512745	1
Nitrogen metabolism	3491601	1
Tropane, piperidine and pyridine alkaloid biosynthesis	3491601	1
Novobiocin biosynthesis	4000758	1
Wnt signaling pathway	588768	1
Notch signaling pathway	588768	1
Pathways in cancer	588768	1
Chronic myeloid leukaemia	588768	1
Steroid biosynthesis	3277060	1
Biosynthesis of terpenoids and steroids	3277060 3286309	2

Table 4-10 KEGG summary for the Post-TE removal assembly.

KEGG pathways that share the same name but are linked to different organisms are grouped together.

KEGG pathway	IDs	Number of unique IDs assigned to it
Alanine, aspartate and glutamate metabolism	4000758	1
Cysteine and methionine metabolism	3274554 4000758 384445	3
Arginine and proline metabolism	4000758	1
Phenylalanine metabolism	4000758 3491601 3291284 384445	4
Phenylalanine, tyrosine and tryptophan biosynthesis	4000758 384445	2
Novobiocin biosynthesis	4000758	1
Glycine, serine and threonine metabolism	3948583 3498205	2
Fatty acid metabolism	3491824 817215 3503787 3489853 3499627 3493055	6
Ubiquinone and other terpenoid-quinone biosynthesis	3510236 3264871 384445	3
Lysine biosynthesis	3512745	1
Tryptophan metabolism	3291284 3516231	2
Alpha-Linolenic acid metabolism	3266521 3490600	2
Limonene and pinene degradation	3507363 3496243 3500475 3506487 3503663 3502969 3494348 3497859	8
Diterpenoid biosynthesis	3499037 3493541 3517679 3509334	4
Brassinosteroid biosynthesis	3499627 3271088 3501700 3286309	4
Carotenoid biosynthesis	3284113 3286008 3493449	3
Nitrogen metabolism	3491601	1
Phenylpropanoid biosynthesis	3510236 3511325 3491601 3264871	4
Flavonoid biosynthesis	3497234 2941875	2
Stilbenoid, diarylheptanoid and gingerol biosynthesis	3507363 3496243 3500475 3506487 3503663 3502969 3494348 3497859	8
Tropane, piperidine and pyridine alkaloid biosynthesis	3491601	1
Biosynthesis of phenylpropanoids	3510236 3511325 3497234 3491601 3264871 2941875	6
Biosynthesis of terpenoids and steroids	3277060 3271088 3286309 3499037 3509334	5
Biosynthesis of alkaloids derived from shikimate pathway	3510236 3491601 3264871 3291284	4
Biosynthesis of alkaloids derived from ornithine, lysine and nicotinic acid	3512745 3491601	2
Biosynthesis of plant hormones	3274554 3266521 3490600 3277060 3286309 3271088 3499037 3491601 3509334	9
Endocytosis	3512745	1
Wnt signaling pathway	588768	1
Notch signaling pathway	588768	1

Pathways in cancer	588768	1
Chronic myeloid leukaemia	588768	1
Glycolysis / Gluconeogenesis	5636578 825753	2
Pyruvate metabolism	5636578 825753	2
Propanoate metabolism	5636578 825753	2
PPAR signaling pathway	817215	1
Adipocytokine signaling pathway	817215	1
Steroid biosynthesis	3277060	1
Histidine metabolism	3291284	1
Flavone and flavonol biosynthesis	3284426	1
Isoquinoline alkaloid biosynthesis	3291284	1
Glyoxylate and dicarboxylate metabolism	3218943 3929207	2
Tyrosine metabolism	4000758 3291284 384445	3
Arachidonic acid metabolism	615542 597566	2
Circadian rhythm	2941875	1

4.3.3.3 EC annotation of GOI daffodil transcripts

As well as the KEGG results, the EC assignments were also used within the DAVID enrichment analysis and so reflect the same results. The most represented enzyme type is cytochrome P450s in all three assemblies (11 DAVID IDs within the 454 and 28 in the post and pre-TE), the most abundant being from the CYP71 clan. Further analysis showed that three DAVID IDs linked to *Catharanthus roseus* were also CYP450s (UniProt IDs Q05047 secologanin synthase, P98183 tabersonine 16-hydroxylase and P48522 Trans-cinnamate-4-monooxygenase). Enzymes already linked to galanthamine biosynthesis are represented in all three assemblies. They include PAL (3 DAVID IDs for the 454, 6 for both Original and Post-TE) and a tyrosine decarboxylase can be seen in both Original (2 IDs) and Post-TE (3 IDs) Trinity assemblies but not the 454. Finally, a transcript linked to NCS can be seen in all three assemblies.

4.3.4 The determination of non-synonymous SNPs within the putative GOI transcripts - Prediction of ORFs and non-synonymous SNPs, confirmation of transcripts via PCR and Sanger sequencing confirmation of SNPs.

(*For ease of reference this section is arranged by transcript, with all relevant information for each transcript together on one data sheet. One transcript from the 454 and one from the post transposon removal Trinity data is shown for reference. The whole data set for both 454 and the post TE Trinity assemblies can be seen in the data_sheets folder of the appendix disc. A summary is given in table 4.17).*

The information on each transcript data sheet encompasses the original fasta sequence of the transcript for reference. The BLASTx results from the GOI database search is shown as it was used to predict the ORF for SNP discovery between Andrew's Choice and Carlton. The original UniProt BLAST hit from the annotation pipeline (sections 2.3.4.5 and 3.4.2.15) is given as it may differ from the hit from the GOI but may have higher homology or show the same class of enzyme. All GO terms linked to the transcript as well as the EC number are shown to since they infer function. Finally the non-synonymous SNPs found between the two varieties are shown as an amino acid alignment with the SNP location, predicted amino acid change and whether sequence was confirmed via PCR are included.

4.3.4.1 4.4.5.1 HDA57HA01AOVJD

Fasta sequence:

```
> HDA57HA01AOVJD length=456
ACGTACACACTGTAATCTTAGAGATCATTAAGTCTACTTGCTCCTTCCGAATATGCCCAAAAGATTGCACCAATTTGGAACTCAA
GAGATGAGCAACACCAAGCTTCCTGGCCTGTCTCCAGCCCTCACCATAGGGTGACAATTGGGAATGGGTGTTTCCGTAGCACAGCA
GCCCGGCCATCTTCGAAGATGGCCTGCCGGCGAAGATGAGATCCTGAGTTTTCAGTATCTCTCGAGCCATCTCTGGTGATGAGATG
ATGAGGGTTGGGACGCAGCCTAGTTGAAGAAGCACCAGACCAGGTGATCCATACTTGTTGGATAATGCACGCAGAGAGCGAGGAG
AGAGGGAGCCGTCGAGCTGGTGGAGGTTGCCTATGATTGGAAGCTTTGGTGGAGATGGTGGGAGGACCTTTGCATCCCAAGATTT
CCTCATGCTTTTTGCAAGAGTGTGTGCGT
```

BLASTx result against the predicted gene database:

```
tr|O64901|O64901_ESCCA (S)-N-methylcoclaurine 3'-hydroxylase
OS=Eschscholzia californica GN=CYP82B1 PE=2 SV=1
Length = 560
Score = 73.9 bits (180), Expect = 3e-16
Identities = 41/125 (32%), Positives = 68/125 (54%), Gaps = 2/125 (1%)
Frame = -3

Query: 403   PPSPPKLPIIGNLHQLDGSLSP--RSLRALSNKYGSPGLVLLQLGCVPTLIISSPEMARE 230
          P +    PI+G+L QL GS  P  R L  +++K+G   + +++ G  PTL++S+ EMA+E
Subject: 44  PEAAGSWPIVGHLPQLVGSGKPLFRVLGDMADKFGP--IFMVRFGVYPTLVVSTWEMAKE 101

Query: 229    ILKTQDLIFAGRPSSKMAGLLCYGNTHSQLSPYGEGWRQARKLGVAHLLSSKLVQSFGHI 50
            + D   A RP S  +  + Y +     S YG  WR+ RK+   HLLS + ++   H+
Subject: 102  CFTSNDKFLASRPPSAASSYMTYDHAMFGFSFYGPYWREIRKISTLHLLSHRRLELLKHV 161

Query: 49    RKEQV 35
          ++
Subject: 162 PHTEI 166
```

UniProt ID from BLAST search against whole UniProt database: P24465

EC number: 1.14

GO IDs: GO:0009835- fruit ripening

GO:0055114- oxidation-reduction process

GO:0004497- monooxygenase activity

GO:0005506- iron ion binding

GO:0016491- oxidoreductase activity

GO:0016705-oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen

GO:0020037- heme binding

GO:0046872- metal ion binding

GO:0005783- endoplasmic reticulum

GO:0005789- endoplasmic reticulum membrane

GO:0016020- membrane

GO:0016021- integral component of membrane

GO:0031090- organelle membrane

GO:0043231- intracellular membrane-bounded organelle

Non-synonymous SNPs between *Narcissus pseudonarcissus var* Carlton and var Andrew’s Choice:

```
Carlton:        PPSPPKLPIIGNLHQLDGSLSPRSLRALSNKYGSPGLVLLQLGCVPTLIISSPEMAREIL

Andrew's choice: PPSPPKLPIIGNLHQLDGSLSPRSLRALSNKYGSPGLVLLQLGCVPTLIISSPEMAREIL

Carlton:        KTQDLIFAGRPSSKMAGLLCYGNTHSQLSPYGEGWRQARKLGVAHLLSSKLVQSFGHIRK
```

Andrew's choice: KTQDLIFAGRPSSKMAGLLCYGNTHSQLSPYGEGWKQARKLGVAHLLSSKLVQSFGHIRK

Carlton: EQV

Andrew's choice: EQV

Nucleotide changes that caused the protein changes:

Position in nucleotide sequence	Reference nucleotide seen in Carlton	Alternative nucleotide seen in Andrew's Choice	Number of Andrew's choice reads same as reference	Number of Andrew's choice reads with alternative nucleotide	Protein change predicted
117	C	T	5	112	R to K

Confirmation of transcript and SNP:

The Transcript was confirmed in both species via RT-PCR and the polymorphism at position 117 was confirmed using Sanger sequencing.

4.3.4.2 4.4.5.2 Daff88927

Fasta sequence:

```
>Daff88927
AGACAGCAGTGTGAAGGTTTCTTTCAAGAGGCCTCGGTTTCAAGGCTAATTTAGATTAAGAAGATCAGAACCAAAATAGATATCA
GGAAAGAAGAAAAAAAGGAAACAAATAGAGTGACTATGAAATTTATATTCAGTTGGTTTAATTCATCTCTGCTTCCTCTCCTCTT
GTTTTTATCCCTCATATTTCTAATCATCAGAAGGCAAATCTCTAAAAATATCAAGCTTCCACCTTCTCCTCCAAAGCTTCCTTTCA
TTGGAAACCTGCACCAACTCCTAAGCAGCTCACTACCCCATCATTCACTCCATGCTCTTTCCAAGAAGTATGGCCCCCTCATGCTT
CTTCAACTTGGTCAGATTCCGACACTTGTAGTCTCATCTCCATATTTTGCCCAAGAAATCCTGAGAACCCATGATGCAGTATTTGC
AAGCAGGCCTTCTAACAAAGCTGCTAGAATTCTGTCATATGGAGGCAGTGACATAACTTTTGCACCTTATGGCGAATATTGGAAG
CAATTGAGAAAGCTTTGCGTTAACCACCTCTTGGGTCCGAAATTGGTGCCATCTTTCCGACGAGTGCGAGAAGAGGAAGTGGCAT
TTATGATTAACGAGATTTCAACAACAAGTTTGTCAACGGGTTCTATAGATCTGACCAAAGTTTTGAACCTGTTCACCAACAACGT
ATTATGTAGAACTGTACTGGGAAAATCATATAGAGGAGAAGAGAAGAATAAAATCTTCTGCGAACTGACGGAGGAGGTTGGTAT
CCTTTTAGGGAAATTATGTGTTGCGGATTACTTTCCTTCCCTCGGATGGTTAGACATGTTTACGGGATTTGAAAGGAGGGCCAGA
AAATGTTCTAAGAGACGGGGTGCTGTTCTTGATGAAGTAATTGATGATTATGTGAGAAATATTAAAGATCCGGATTATAAATCC
GAAAACAGACATTTTGTGGAAGCTCTGCTTGATCCTCGGAATGATACTAGTGCAGAATTTCCAGTAAACAGAGAGATGATCAAGA
TACTCATACAGGATATGATCGGAGCAGGGACCGACACGTCATTTGTAACCTTAGAATGGGGAATGGCAGAGCTTATTCGCAACCC
AAAAGTGATGAAAAAACTGCAAGACGAAACAAGAGGTGGTGCCAACAAAGAGTATTCCATGATTATGGAGGAAGATTTAAGCAA
AATGACCTACCTGAAAGCTGTTATAAAGGAGATCCTGAGATTGCATCCTCCGGCTCCTTTGATGATTCCTCATGAATCAATGGAA
GATTGCAAGATACAGGGCTACGAAATTCCAAAGAAAACAAGAGTTGTTGTCAATGCTTGGGCAATCGGTAGGGACCCTGGATTTT
GGGATGCACCAGAGGAGTTCCGACCTGAGAGATTTCTGAATAGCTCGGTTGATTTCGCTGGGCATGATTTCGAATTCATTCCATT
TGGAGCAGGTCGAAGAATTTGTCCTGGAATGCAGTTTGCAATAGCGACTTTGAAGCTTGCTTTAGCAAATCTCGTCCTTCGGTTT
GACTGGAAATTATCTGAAGAGATTGAATCTAAGAGTATAGACATGAGTGAAGTTCCCGGATTGACCAGTCATAAGAAGGAGAAG
CTAAACTTGGTCGCCAAACCAGCTTTTATCATGTAAATTCAGGGGATTATAGTCTGAGAGCATATGTGTTGTCTATATTAATGTA
TTTCTATTTGTGTTGAACTCTAGGAATTTGAATTTGATATTTTCTACATAAGATTGTTGATGGATCTTGTGGTGACAGAGAAGAA
AGTAATTTTGTGTTGTTCTTCCAGTGCCAGGGAAAG
```

BLASTx result against the predicted gene database:

```
tr|C3SBT0|C3SBT0_PAPBR Putative (S)-N-methylcoclaurine 3'-hydroxylase(Fragment)
OS=Papaver bracteatum
PE=2
SV=1
Length = 486
Score =  276 bits (706), Expect = 2e-76
Identities = 168/474 (35%), Positives = 245/474 (51%), Gaps = 5/474 (1%)
Frame = +1

Query: 226    LPPSPPKLPFIGNXXXXXXXXXXXXXXXXXXKKYGPLMLLQLGQIPTLVVSSPYFAQEIL 405
           LPP P   P +GN                  K YG L  L+LG    +V S+P  A EIL
Subject: 28   LPPGPKPWPIVGNLLQLGEKPHSQFAELA--KTYGDLFTLKLGSETVVVASTPLAASEIL 85

Query: 406    RTHDAVFASRPSNKAARILSYGGSDITFAPYGEYWKQLRKLCVNHLLGPKLVPSFRRVRE 585
           +THD V + R   ++ R+  +  + I ++   E WK+LRK+C   L   K++ S   +RE
Subject: 86   KTHDRVLSGRYVFQSFRVKEHVENSIVWSECNETWKKLRKVCRAELFTQKMIESQAEIRE 145

Query: 586    EEVAFMINEIST---TSLSTGSIDLTKVLNLFTNNVLCRTV--LGKSYRGEEKNKIFCEL 750
            +   M+  +    ++    +    ++N+F N +  ++ +  LG    G  + K
Subject: 146  SKAMEMVEFLKRNQGSEVKIVEVVFGTLVNIFGNLIFSQNIFKLGDESSGSVEMKEHLWR 205

Query: 751    TEEVGILLGKLCVADYFPSLGWLDMFTGFERRARKCSKRRGAVLDEVIDDYVRNIKDPDY 930
            E+G       ADYFP LG  D+F G  +    C +   +V   ++ +  R I
Subject: 206  MLELG---NSTNPADYFPFLGRFDLF-GQRKDVADCLQGIYSVWGAMLKE--RKIAKLHN 259

Query: 931    KSENRHFVEALLDPRNDTSAEFPVNREMIKILIQDMIGAGTDTSFVTLEWGMAELIRNPK 1110
           S+   FVE LLD   D        + I  L+ ++ GAGT+TS  T+EW ++EL +NP+
Subject: 260  NSKKNDFVEILLDSGLDD--------QQINALLMEIFGAGTETSASTIEWALSELTKNPE 311

Query: 1111   VMKKLQDETRGGANKEYSMIMEEDLSKMTYLKAVIKEILRLHPPAPLMIPHESMEDCKIQ 1290
          V   ++ E      K    + E D+  M YL+A +KE LRLHP  PL++P  ++E CK+
Subject: 312  VTANMRSELLSVVGKR--PVKESDIPNMPYLQAFVKETLRLHPATPLLLPRRALETCKVL 369

Query: 1291   GYEIPKKTRVVVNAWAIGRDPGFWDAPEEFRPERFLNSSVDFAGHDFEFIPFGAGRRICP 1470
           Y IPK+ +++VNAW IGRDP  W  P +F PERFLNSS+DF G+DFE IPFGAGRRICP
Subject: 370  NYTIPKECQIMVNAWGIGRDPKRWTDPLKFAPERFLNSSIDFKGNDFELIPFGAGRRICP 429

Query: 1471   GMQFAIATLKLALANLVLRFDWKLSEEIESKSIDMSEVPGLTSHKKEKLNLVAK 1632
          G+  A   + L +  LV  FDW L + ++   + M E  GLT  K+  L +V K
Subject: 430  GVPLATQFISLIVPTLVQNFDWGLPKGMDPSQLIMEEKFGLTLQKEPPLYIVPK 483
```

UniProt ID from BLAST search against whole UniProt database: P24465
EC number: 1.14.-.-
GO IDs: GO:0009835- fruit ripening
GO:0055114- oxidation-reduction process
GO:0004497- monooxygenase activity
GO:0005506- iron ion binding
GO:0016491- oxidoreductase activity
GO:0016705-oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0020037- heme binding
GO:0046872- metal ion binding
GO:0005783- endoplasmic reticulum
GO:0005789- endoplasmic reticulum membrane
GO:0016020- membrane
GO:0016021- integral component of membrane
GO:0031090- organelle membrane
GO:0043231- intracellular membrane-bounded organelle

Non-synonymous SNPs between *Narcissus pseudonarcissus var* Carlton and var Andrew's Choice:

Carlton: LPPSPPKLPFIGNLHQLLSSSLPHHSLHALSKKYGPLMLLQLGQIPTLVVSSPYFAQEIL

Andrew's choice: LPPSPPNLPFIGNLHQLLSSSLPHHSLHALSKKYGPLMLLQLGQIPTLVVSSPYFAQEIL

Carlton: RTHDAVFASRPSNKAARILSYGGSDITFAPYGEYWKQLRKLCVNHLLGPKLVPSFRRVRE

Andrew's choice: RTHDAVFASRPSNKAARILSYGGSDITFAPYGEYWKQLRKLCVNHLLGPKLVPSFRRVRE

Carlton: EEVAFMINEISTTSLSTGSIDLTKVLNLFTNNVLCRTVLGKSYRGEEKNKIFCELTEEVG

Andrew's choice: EEVAFMINEISTTSLSTGSIDLTKVLNLFTNNVLCRTVLGKSYRGEEKNKIFCELTEEVG

Carlton: ILLGKLCVADYFPSLGWLDMFTGFERRARKCSKRRGAVLDEVIDDYVRNIKDPDYKSENR

Andrew's choice: ILLGKLCVADYFPSLGWLDMFTGFERRARKCSKRRGAVLDEVIDDYVRNIKDPDYKSENR

Carlton: HFVEALLDPRNDTSAEFPVNREMIKILIQDMIGAGTDTSFVTLEWGMAELIRNPKVMKKL

Andrew's choice: HFVEALLDPRNDTSAEFPVNREMIKILIQDMIGAGTDTSFVTLEWGMAELIRNPKVMKKL

Carlton: QDETRGGANKEYSMIMEEDLSKMTYLKAVIKEILRLHPPAPLMIPHESMEDCKIQGYEIP

Andrew's choice: QDETRGGANKEYSMIMEEDLSKMTYLKAVIKEILRLHPPAPLMIPHESMEDCKIQGYEIP

Carlton: KKTRVVVNAWAIGRDPGFWDAPEEFRPERFLNSSVDFAGHDFEFIPFGAGRRICPGMQFA

Andrew's choice: KKTRVVVNAWAIGRDPGFWDAPEEFRPERFLNSSVDFAGHDFEFIPFGAGRRICPGMQFA

Carlton: IATLKLALANLVLRFDWKLSEEIESKSIDMSEVPGLTSHKKEKLNLVAK

Andrew's choice: IATLKLALANLVLRFDWKLSEEIESKSIDMSEVPGLTSHKKEKLNLVAK

Nucleotide changes that caused the protein changes:

Position in nucleotide sequence	Reference nucleotide seen in Carlton	Alternative nucleotide seen in Andrew's Choice	Number of Andrew's choice reads same as reference	Number of Andrew's choice reads with alternative nucleotide	Protein change predicted
246	G	C	171	176	K to N

Confirmation of transcript and SNP:

Transcript was confirmed in both varieties via PCR. The SNP was confirmed via Sanger sequencing. With both Andrew's Choice sequences assigned C and Carlton forward assigned as G.

4.3.4.3 Non-synonymous SNP verification

Any of the transcripts that were predicted to have non-synonymous SNPs using the ORF predicted by the BLASTx search against the GOI were analysed via Sanger sequencing for confirmation. The results are shown in table 4.17.

Table 4-11 Summary of SNP mining of GOI transcripts.

SNPs were considered confirmed if Andrew's Choice showed the alternative nucleotide only or if both nucleotides were present in Andrew's Choice. It was unconfirmed if the reference was only seen in both or if both the reference and the alternative were seen in Carlton.

<table>
<tr><th>Transcript ID</th><th>UniProt ID hit</th><th>Name</th><th>SNP confirmed via Sanger sequencing</th></tr>
<tr><td>Contig 01152</td><td>tr|Q7XB10|Q7XB10_PAPSO</td><td>S-adenosyl-L-methionine: 3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase 2</td><td>Unconfirmed</td></tr>
<tr><td>contig01153</td><td>sp|Q9LEL6|6OMT_COPJA</td><td>(RS)-norcoclaurine 6-O-methyltransferase</td><td>confirmed</td></tr>
<tr><td>Contig02456</td><td>sp|Q39224|SRG1_ARATH</td><td>Protein SRG1</td><td>confirmed (both nucleotides present)</td></tr>
<tr><td>HDA57HA01BEL8O</td><td>O064900|C80B2_ESCCA</td><td>CYP80B2</td><td>no Sanger sequencing</td></tr>
<tr><td>HDA57HA01OVJD</td><td>tr|O064901|O064901_ESCCA</td><td>S)-N-methylcoclaurine 3'-hydroxylase</td><td>Confirmed</td></tr>
<tr><td>Daff62300</td><td>sp|Q39224|SRG1_ARATH</td><td>Protein SRG1</td><td>confirmed (both nucleotides present)</td></tr>
<tr><td>DAff74804</td><td>tr|Q7XB10|Q7XB10_PAPSO</td><td>S-adenosyl-L-methionine:3'-hydroxy-N-methylcoclaurine4'-O-methyltransferase 2</td><td>confirmed (both nucleotides present)</td></tr>
<tr><td>Daff106212</td><td>sp|Q9LEL6|6OMT_COPJA</td><td>(RS)-norcoclaurine 6-O-methyltransferase</td><td>confirmed</td></tr>
<tr><td>Daff88927</td><td>tr|C3SBT0|C3SBT0_PAPBR</td><td>Putative (S)-N-methylcoclaurine 3'-hydroxylase</td><td>Confirmed</td></tr>
<tr><td>comp97312_c0_seq1</td><td>sp|O064899|C80B1_ESCCA</td><td>(S)-N-methylcoclaurine 3'-hydroxylase isozyme 1</td><td>confirmed (both nucleotides present)</td></tr>
<tr><td>comp99544_c0_seq1</td><td>tr|O064901|O064901_ESCCA</td><td>(S)-N-methylcoclaurine 3'-hydroxylase</td><td>confirmed</td></tr>
<tr><td>comp100406_c1_seq2</td><td>tr|Q6WUC2|Q6WUC2_PAPSO</td><td>(R,S)-reticuline 7-O-methyltransferase</td><td>confirmed</td></tr>
<tr><td>comp100760_c0_seq2</td><td>tr|Q9FQY6|Q9FQY6_CAPAN</td><td>Cinnamic acid 4-hydroxylase</td><td>confirmed (both nucleotides present)</td></tr>
</table>

4.3.5 The determination of significant differences in transcript levels for transcripts selected from the GOI database

4.3.5.1 BitSeq

The BitSeq analysis of transcripts from the GOI database search resulted in a total of 13 transcripts being predicted to show differential expression between the two daffodil cultivars (see a summary in table 4.18). Six were at a significantly lower level in Andrew's Choice, comprising 4 from the total of 77 transcripts within the 454 data and 2 from the total of 201 transcripts within the Post-TE removal Trinity data. The number of transcripts predicted to show significantly higher expression in Andrew’s Choice was 7 with 3 from the 454 and 4 from the Illumina data. Full results can be seen on appendix disc.

4.3.5.2 Confirmation of levels via comparative qPCR

Of the 13 transcripts with an indication of differential expression using BitSeq, only 9, were carried forward to qPCR for confirmation of expression differences. The results are shown in table 4.18 below. This indicated that the expression differences proposed via BitSeq were confirmed for seven of the 9 transcripts. This is shown in figure 4-23, the two transcripts that showed contradicting qPCR results to the BitSeq predictions are HDA57HA01AW38A and HDA57HA01AK3FX from the 454 data. This difference could be linked to the lower coverage seen for the 454 data, both transcripts are singletons and so the sequencing data may not have been of high enough quality or coverage to accurately predict a transcript level difference. The qPCR was run on three biological samples with three technical replicates and so is likely to give a better insight into the transcript level differences. Both singletons showed relatively low fold changes (<2) and so it may be that these transcripts do not show a significant difference between the two cultivars. Overall the qPCR results confirmed the BitSeq results suggesting BitSeq as a suitable method for predicting transcript level differences between cultivars lacking replicated sequence data.

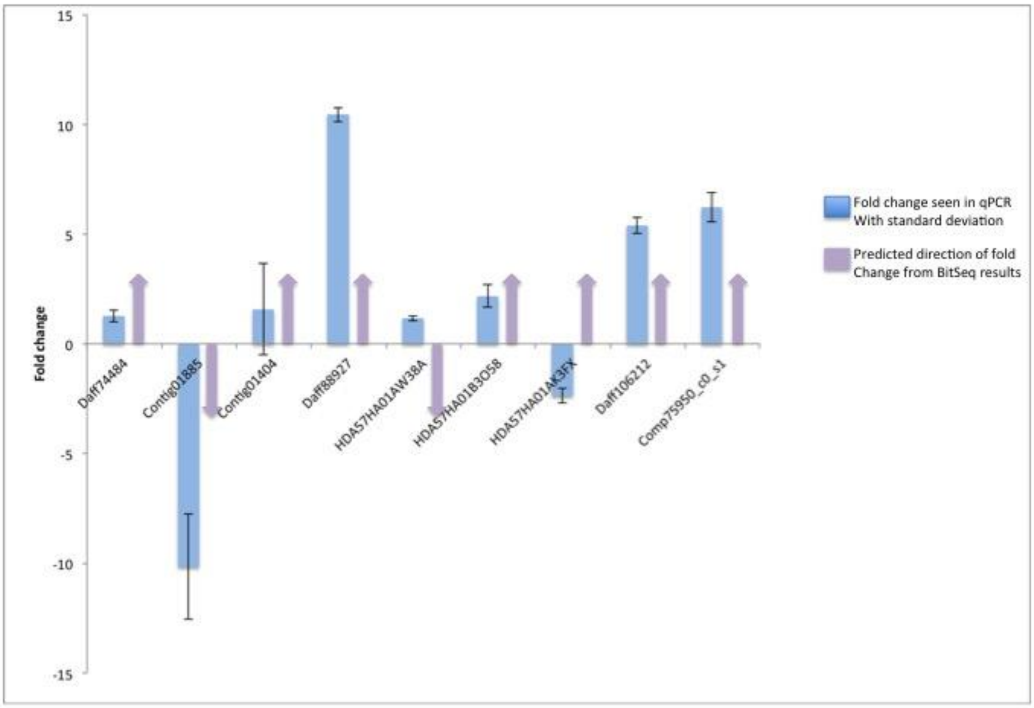


Figure 4-23 Graph to show fold change in transcript level seen in Andrew's Choice compared to Carlton

The graph shows the actual fold changes seen via qPCR using actin as a control and the predicted direction of fold change from the BitSeq results. A PPLR value of over 0.95 suggesting a significant positive fold change and below 0.05 as a significant negative fold change.

As an example, the trace produced for Daff88927 is shown in figure 4-23. For simplicity one biological replicate is pulled out and shown in the trace in figure 4-24. It is clear from both traces that a clear difference between the two varieties is seen with Andrew's Choice showing a fold change of 10.44 (s.d. 0.32) corresponding to a PPLR value of 0.978. It is also clear that actin is very similar in both varieties and is therefore a suitable candidate for a control gene.

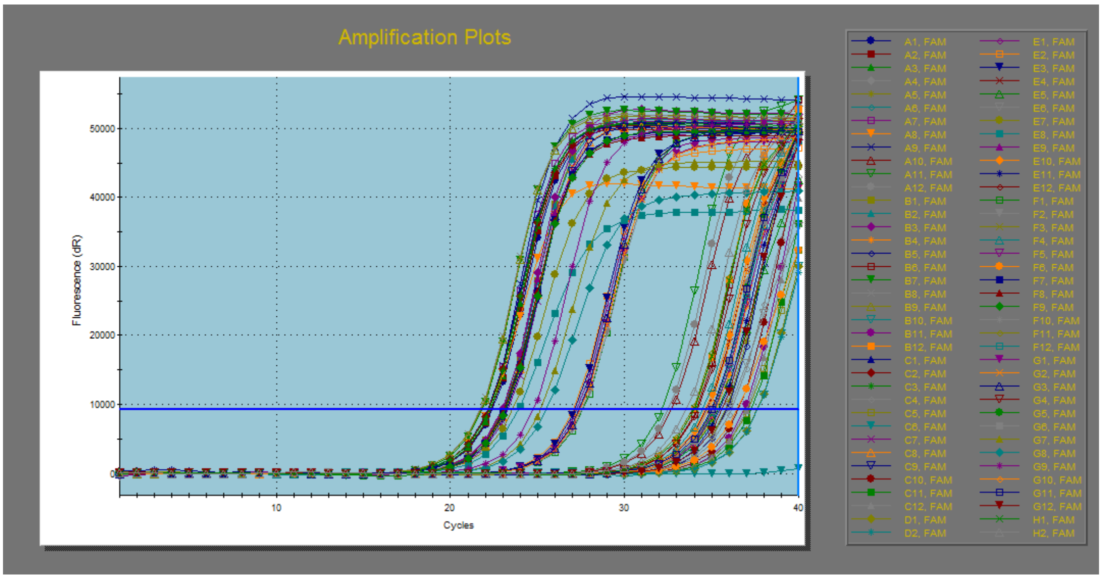


Figure 4-24 Amplification trace for Daff88927

The trace clearly shows grouping for both the Andrews Choice (E7-F9, F7-F9 and G7-G9) and Carlton samples (A1-A3, B1-B3 and C1-C3) as well as the actin samples for both varieties (Andrews Choice A7-A9, B7-B9 and C7-C9, Carlton E1-E3, F1-F3 and G1-G3). The actin traces are all similar suggesting it is a good candidate for a control gene, the Andrews Choice traces are all to the left of the actin and the Carlton to the right showing a clear difference in expression. All negative controls are close to the base line suggesting no interference from genomic DNA or contaminants.

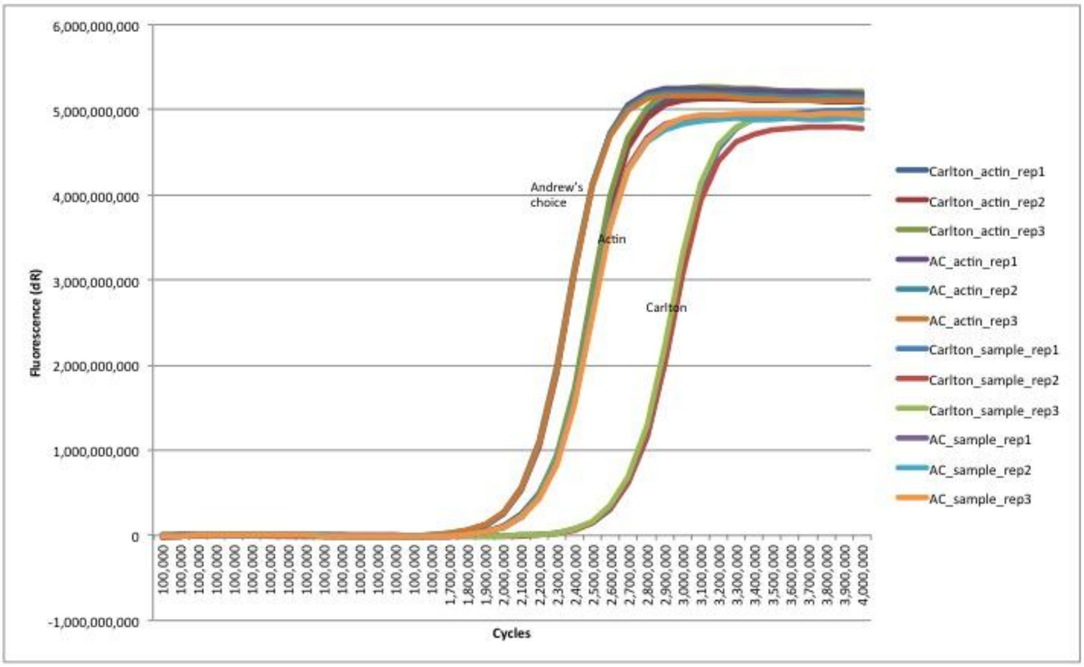


Figure 4-25 Simplified Trace for Daff88927 showing one biological replication

The simplified trace shows the three technical replicates for one biological replicate. It clearly shows the difference between the two varieties for gene expression of Daff88927 compared to actin. The traces are labeled Actin, Carlton and Andrew's Choice to show the groupings.

Table 4-12 Transcript level result summary.

Shows observed fold change via qPCR, BitSeq PPLR values and the results of both the UniProt wide BLASTx search from the annotation pipeline (sections 2.3.4.5 and 3.4.2.15) and the BLASTx search against GOI. Both results are shown for comparison and further evidence for predicted function within galanthamine biosynthesis.

Transcript ID	qPCR Fold Change	qPCR Standard deviation	BitSeq PPLR value	Whole UniProt BLAST result	Description of UniProt ID	Homology Shown to UniProt result (%)	GOI database BLAST result	Description of GOI ID	Homology shown to GOI result (%)
Daff74484	1.26	0.27	0.978	Q8GSN1	MOMT_CATRO Myricetin O-methyltransferase	40.89	Q9LEL6	6OMT_COPJA (RS)-norcoclaurine 6-O-methyltransferase	37.70
Contig01885	-10.16	2.40	0.016	B6YWH0	GYAR_THEON Glyoxylate reductase	45.39	Q65CJ7	Q65CJ7_SOLSC Hydroxyphenylpyruvate reductase	72.85
Contig01404	1.57	2.08	0.996	Q05736	PR1_ASPOF Pathogenesis-related protein 1	62.00	C3SBV5	C3SBV5_9MAGN Norcoclaurine synthase	33.70
Daff88927	10.44	0.32	0.978	P24465	C71A1_PERAE Cytochrome P450 71A1	48.63	C3SBT0	C3SBT0_PAPBR Putative (S)-N-methylcoclaurine 3'-hydroxylase	36.29
HDA57HA01AW38A	1.16	0.11	0.05	O04130	SERA_ARATH D-3-phosphoglycerate dehydrogenase 2	78.36	Q65CJ7	Q65CJ7_SOLSC Hydroxyphenylpyruvate reductase	33.1
HDA57HA01B3O58	2.18	0.51	0.984	P28002	COMT1_MEDSA Caffeic acid 3-O-methlytransferase	57.36	G3FDY0	3FDY0_SALMI Caffeic acid O-methyltransferase	60.47
HDA57HA01AK3FX	-2.37	0.33	0.972	P37119	C71A3_SOLME Cytochrome P450 71A3	42.00	O64900	C80B2_ESCCA (S)-N-methylcoclaurine 3'-hydroxylase isozyme 2/Cytochrome P450 80B2	29.29
Daff106212	5.39	0.37	0.98	Q8GSN1	MOMT_CATRO Myricetin-methyltransferase	52.11	Q9LEL6	6OMT_COPJA (RS)-norcoclaurine 6-O-methyltransferase	44.44
Comp75950_c0_s1	6.22	0.67	0.98	P47787	THAS_PIG Thromboxane-A-synthases CYP5A1	24.47	Q9FQY6	Q9FQY6_CAPAN Cinnamic acid 4-hydroxylase	22.09

4.4 Discussion

4.4.1 GO annotation and categorization of the whole transcriptome

The GO ID assignments and subsequent categorization using the UniProt gene ontology option within its retrieval program provided a rapid method of putative functional annotation of the transcriptome. This representation of a transcriptome has been used in many other non-model plant projects but requires further information to infer functionality with confidence [5,93,97,110].

There are no GO annotations for species closely related to daffodils but it is important that the suggested functionalities are further scrutinized since paralogs may have acquired different functionality, especially since more distantly related species have been used to confer UniProt and GO IDs [189]. An example of this is seen with NCS in *Coptis japonica* that is involved in BIA biosynthesis but shows high protein homology (42%) with the SRG1 protein in *Arabidopsis thaliana* that totally lacks norcoclaurine synthase activity but may be involved in amine or aldehyde detoxification [268].

The results of the GO analysis in this project contained both expected and unexpected annotations. All three assemblies show similar profiles with the most highly represented molecular functions being binding and catalytic activity in all three. As the plant tissue used was basal plate containing meristematic tissue harvested at a time of active growth and when the plant was known to have its highest level of alkaloids (Dr X. Chang, personal communication) it would be reasonable to assume that substantial catalytic activity was occurring within the samples.

Within the biological processes category all three assemblies showed a high number of IDs associated with single-organism processes (a process that only involves one organism). Closer inspection of the UniProt IDs associated with this term shows that many of these IDs are from plants, (in the 454 data 1062 of the 2710 IDs were linked to *Arabidopsis thaliana*, 356 to rice and 41 to tobacco) but rat IDs accounted for 1125 and a few associated with single-celled organisms such as *E. coli* (7) and *Pseudomonas* (4). This is a further example of the need to complement GO functionality predictions with other methods of analysis as this GO term encompasses a broad spectrum of processes and so cannot confer any specific functionality. It has 28 child terms, the biggest being single-organism developmental processes that itself has 446 child terms.

GO terms are often quite broad and so without going further down the layers of categorization it is impossible to determine function from broad terms such as single-organism processes. Also there is a substantial knowledgebase for model organisms such as *Arabidopsis* and rat and it is possible that these could cause bias and lead to incorrect annotation.

4.4.2 BLASTx results

A literature search (sections 4.1.3 and 4.2.2) was used to produce a list of genes already known to be involved in alkaloid biosynthesis, as well as homologs of such genes from non-alkaloid producing plants and this was used to search for orthologs in daffodils. The BLASTx search against this database (section 4.2.3) resulted in the prediction of over 600 transcripts from the three assemblies. This list was then used as the gene of interest list for the DAVID enrichment analysis as well as for further investigation into putative genes via SNP and DE studies. This sort of comparison, against transcripts strongly linked to the desired function (alkaloid production in this case), frequently from experimental data, is an important further screen to determine putative transcripts with these functions in daffodil.

4.4.3 DAVID analysis, EC and KEGG annotation

By focusing in on the transcripts predicted via the BLASTx search against the GOI database it was possible to carry out functional profiling to answer questions relating to the enrichment of specific pathways such as alkaloid production [189].

The DAVID program predicted enrichment scores as well as annotating transcripts with corresponding KEGG and EC numbers (user defined options in analysis). The results showed both expected and unexpected results (as indicated in section 4.33). Enrichment of P450s linked to oxidoreductase and secondary metabolism would be expected in samples of this type when looking at specific genes thought to be involved in alkaloid production. However, even when looking at clusters with enrichment score as high as those seen with the P450s (14.74 for the 454 data, 38.82 and 35.13 for the Original and Post-TE removal assemblies) it is important to also look at the P-values (EASE scores) and the fold enrichment scores [193,291]. More specific terms that may only have a few genes associated to them may appear enriched when they are not [197] and likewise terms with high numbers of genes can be missed.

The daffodil data suggest that P450s, methyltransferases, phenylpropanoid biosynthesis and PAL are all enriched as they have a high percentage of representation in the GOI list, low p-values, high enrichment scores and FE scores above 1.3 (ranging from12 to 74). Enrichment analysis of transcripts of particular interest such as those annotated as P450, PSRs, PAL and tyrosine decarboxylase is all discussed in sections 4.4.5.1to 4.4.5.7.

The DAVID results showed enrichment for these genes in the Carlton variety of daffodil. However, one aim of this project is to compare Carlton with Andrew's Choice as a further strategy to identify putative genes involved in galanthamine production and so the GOI was taken forward to SNP and transcript level analysis to discover whether any of the enriched clusters also had differences in these characters (see section 4.3.4 and 4.3.5).

As already discussed in section 4.1.1.2, KEGG has similar pitfalls in its ability to accurately predict function of orthologs in non-model organisms due to broad or very specific pathways that may or may not be present [184]. This can be seen in the daffodil data with the ID 588768 (UniProt Q0VCQ1) association to the "pathways in cancer" and "chronic myeloid leukemia" pathway terms. Further inspection of this ID show that it is a C-terminal binding protein in *Bos taurus* (CtBP2), targeting transcriptional regulators in brown fat and neural tissues. This protein is a transcription repressor domain attached to an N-terminal domain of a protein known as RIBEYE. Both CtBP2 and RIBEYE are encoded in the same gene but due to alternative splicing have different functions (CTBP2 is used for transcriptional control, RIBEYE is used for synaptic vesicle transport and exocytosis) but both are derived from a member of the dehydrogenase family [292].

This is an example where inappropriate annotation from sequence similarity searches may predict pathways that are not present in the species of interest (daffodil). However, the daffodil transcript daff66090 shows 36.46% homology to Q0VCQ1 with an e-value of 2e-24 and so could have a very different function. Q0VCQ1 is a member of the d isomer specific 2-hydroxyacid dehydrogenase families that has 84,710 hits on UniProt. Therefore it is very likely that this transcript is not involved in the pathway suggested in KEGG. In other plants, such as Arabidopsis, enzymes of this type are involved in the controlling of the equilibrium between tubular and stacked structure in the Golgi complex among other functions and do not possess the amino acids known to be involved in the dehydrogenase activity [293]. The bovine protein does contain the conserved residues involved in dehydrogenase including the four residues involved in NAD+ binding and the amino acids linked to catalytic function [294].

Another area of difficulty within KEGG annotations is the similarity seen between some terms. For example, the two terms “biosynthesis of phenylpropanoids” and “phenylpropanoid biosynthesis” are very similar, so could be expected to relate to the same transcripts. However, this is not the case. There is a slight difference between these pathways in KEGG. The phenylpropanoid biosynthesis map00940 only involves pathways directly linked to their production from phenylalanine whereas map01061 (biosynthesis of phenylpropanoids) shows a wider view including connected pathways involving lignans and flavonoids.

It is also possible that pathways known to exist in species can also be missed. An example is the case of isoquinoline alkaloid biosynthesis in the daffodil data, since although PAL is known to be involved, and is found in all three assemblies, the term "isoquinoline alkaloid biosynthesis" was only shown in the Illumina data for transcripts linked to TYDC.

4.4.4 Transcript level and SNP analysis

The transcript level and SNP analysis was used to look for differences in transcripts predicted to be involved in galanthamine production and are therefore discussed alongside the discussion of specific gene types (section 4.4.5). The BitSeq analysis produced some results that differ from the corresponding qPCR results and that are worth noting. Both HDA57HA01AW38A and HDA57HA01AK3FX showed opposite results to those predicted. The first was given a PPLR value of 0.05 that should have resulted in lower expression in Andrew's Choice but a 1.16 ± 0.11 fold change increase was seen via qPCR. The second transcript had a PPLR value of 0.972 that should have resulted in a higher level of expression in Andrew's Choice but a 2.37 ± 0.33 fold change decline in expression was observed. PPLR values of 0.05 and 0.095 are the suggested thresholds for significance so it is possible that the levels of these two transcripts are indeed not significantly different between the two cultivars. The qPCR results also suggest only a slight fold change. The qPCR results are from three technical and biological replicates while the BitSeq analysis is from one sequenced sample for each variety. This is an argument for replication in the BitSeq analysis, whenever possible.

4.4.5 Transcripts linked to alkaloid biosynthesis

4.4.5.1 PAL and TYDC

Both PAL and TYDC have been shown to be involved in the first step of galanthamine production as well as other isoquinoline and benzylisoquinoline alkaloids [4,7]. However, although both were found in the Original and post- TE assemblies, and PAL in the 454 assembly, EC identified only one transcript associated with the KEGG term "isoquinoline alkaloid biosynthesis" namely UniProt ID Q7XHL3 from rice. Although rice does not produce alkaloids, the conversion of tyrosine to tyramine is involved in a large number of metabolic pathways and is in no way specific to alkaloid biosynthesis. The Original assembly assigned tyrosine decarboxylase related EC numbers to 2 transcripts in the GOI and 3 were assigned in the Post-TE assembly. The three IDs were Q7XHL3, Q8RY79 (tyrosine decarboxylase 1 from *Arabidopsis thaliana)* and P54769 (tyrosine/DOPA decarboxylase 2 from *Papaver somniferum)*. Although *Arabidopsis thaliana* and rice do not produce isoquinoline alkaloids, *Papaver somniferum* produces BIA but neither BIA or isoquinoline alkaloid biosynthesis pathways were predicted to involve homologs of P54769 via KEGG.

Likewise several transcripts were associated to PAL (3 for 454 and 6 for both Original and Post-TE) none of which were associated in KEGG with alkaloid biosynthesis. The six IDs are shown in table 4.19. A ClustalO alignment of these proteins via UniProt showed a 61.05% identity among the sequences and, as can be seen in the table 4.19, the daffodil transcripts also show close homology to all six PAL sequences (70-90%) (See appendix section 6.9 for ClustalO alignment). Although none of the IDs are linked to organisms that produce isoquinoline alkaloids, they all involve PAL in the conversion of phenylalanine to *trans*-cinnamate and ammonia and so were investigated further through SNP profiling and transcript levels. However none of the transcripts linked to TYDC or PAL from the GOI showed non-synonymous SNPS or transcript level difference between the two daffodil varieties, suggesting that this step is preserved between as both enzymes are important to plant metabolism.

Table 4-13 Combined results from the ClustalO alignment in UniProt with the BLASTx results against the GOI for daffodil transcripts.

UniProt ID	Entry name	Protein name	Organism	Family	Daffodil transcript homology (%) (Post –TM)	E-value	Gene name
P45724	PAL2_ARATH	Phenylalanine ammonia-lyase 2	*Arabidopsis thaliana*	Brassicaceae	73.15	1e-62	PAL2
P45727	PALY_PERAE	Phenylalanine ammonia-lyase	*Persea americana* (avocado)	Lauraceae	86.76	1e-27	PAL
P45729	PAL3_PETCR	Phenylalanine ammonia-lyase 3	*Petroselinum crispum* (parsley)	Apiaceae	90.51	5e-167	PAL3
P45726	PALY_CAMSI	Phenylalanine ammonia-lyase	*Camellia sinensis* (tea)	Theaceae	71.74	2e-29	PAL
Q42609	PALY_BROFI	Phenylalanine ammonia-lyase	Bromheadia finlaysoniana (pale reed orchid)	Orchidaceae	79.72	1e-62	PAL
O64963	PAL1_PRUAV	Phenylalanine ammonia-lyase 1	*Prunus avium* (bird cherry)	Rosaceae	85.62	4e-147	PAL1

4.4.5.2 CYP450s

P450s are predicted to account for up to 1% of plant genome annotations and 5100 P450 associated sequences in plants had been annotated and named by 2011 [295]. The most widely represented clan of P450 in the enrichment analysis was that of CYP71 (see section 4.3.3) This could be anticipated since this clan represents over 50% of all CYPs in higher plants [295]. Members of this clan have been linked to many reactions including metabolism of aromatic and aliphatic amino acid derivatives such as phenylpropanoids, indole derivatives, small isoprenoids and some alkaloids [295]. Five daffodil transcripts showed homology to P450s and also had transcript level differences or non-synonymous SNPS between the two daffodils varieties as discussed below.

4.4.5.3 HDA57HA0AK3FX

This transcript from the 454 data showed 42% homology to a CYP71A3 in the whole UniProt blast search and 20% to the GOI CYP80B2. CYP80 enzymes are responsible for phenyl coupling in other alkaloid biosynthetic pathways such as morphine biosynthesis in *Papaver somniferum* and in *Coptis japonica* and *Eschscholzia californica* [62,217,231]. Both CYP80 and CYP719 enzymes have been shown to be involved in C-C intermolecular coupling and so it is important that not only sequence similarity is taken into account but also evolution and known functions [295]. This transcript could be investigated further for a possible link to the final coupling step in galanthamine production as it showed a transcript level difference (-2.37 ± 0.33) between the two varieties via qPCR (not confirmed via PPLR) indicating a higher level in Carlton (the higher galanthamine producer).

4.4.5.4 Daff88927

This transcript was found in both the Original and post-TM assembly and shows homology to CYP71A1 in the whole UniProt BLASTx search (48.63%) and a P450 NMT in the GOI search (36%). The sequence homology of over 40% and below 50% suggest that it is from the same family but is a member of a different sub family [208]. This transcript showed a 10 fold increase in transcript levels in Andrew's Choice compared to Carlton (10.44 ± 0.32). Evidence of involvement in alkaloid production is supported by the fact that the NMT in *Papaver* *bracteatum* has been showed to be involved in phenol coupling in BIA biosynthesis and so could be involved in the final step of galanthamine biosynthesis [121]. This is a prime candidate for further functionality testing.

4.4.5.5 Comp75950_c0_s1

This transcript was found in only the Post-TE assembly and again showed homology to P450s, specifically to CYP5A1 from pig and a P450 C4H in Capsicum. This transcript shows the second highest fold change seen in the GOI (6.22 ± 0.67) and should be investigated further to look at its functionality compared to C4H which is known to catalyse steps in vanillin biosynthesis in Capsicum and so could be involved in the production of protocatechuic acid in galanthamine biosynthesis [21,245].

4.4.5.6 Comp97312_c0_s1 and Comp99544_c0_s1

These two transcripts from the post-TM assembly showed homology to CYP71A1 and CYP81E1 respectively. Although they did not show any transcript level differences they both had a non-synonymous SNP between the two varieties.

4.4.5.7 Other transcripts of interest

Two other enzymes that could be involved in galanthamine biosynthesis that showed both transcript level differences and SNPs are (R,S) norcoclaurine 6OMT and NCS (Daff106212 and Contig 01404). Both enzymes are involved in BIA biosynthesis and could perform similar roles in daffodils. 6OMT coverts (S)-norcoclaurine to (S)-coclaurine and NCS is a PSR involved in the reaction of 4-HPAA and dopamine to give (S)-norcoclaurine [18,204,227-229]. These could be involved in the sequential conversion of protocatechuic acid and tyramine to norbelladine, and subsequently to 4'-*O*-methylnorbelladine steps within galanthamine biosynthesis. Both show non-synonymous SNPs between the two daffodil varieties. Interestingly, Daff106212 also shows a high fold increase in transcript levels in Andrew's Choice (5.39 ± 0.37) and therefore could be suggested for further investigation linked to metabolite profiling in both varieties as a point in divergence in the alkaloid pathway.

5 Chapter 5 – Overall Conclusions and Future work

5.1 Transcriptomic reference for the daffodil

With the continuing developments in functional genomics specifically towards sequencing and assembling de novo transcriptomes, projects about non-model plants from families with limited to no genomic data are now possible [12,83]. The use of the now out dated 454 platform resulted in a reference transcriptome that could then be used successfully to map Illumina reads from two varieties of daffodil to search for SNP and transcript level differences linked to galanthamine biosynthesis. The Illumina sequencing for the Carlton variety was also successfully assembled, increasing the reference data for daffodils. The 454 assembly resulted in 45,324 transcripts of which 67% were annotated via a BLASTx search pipeline while the Illumina reads produced a total of 165,065 transcripts of which 38% was annotated. Both the 454 and Illumina transcripts have been deposited in the short read archive and will be released along with any publication from this work (http://www.ncbi.nlm.nih.gov/sra).

5.2 The uses of currently available Ontologies and Functional Annotations –bias towards model organisms

As with all aspects of functional genomics, ontologies and databases such as KEGG were developed alongside the data that started being produced for model organisms in the late twentieth century [184,186]. GO started with the model organism databases of Flybase, MGI and SGD, whereas KEGG was initiated for the Human Genome Project [186]. Although intended to be non species-specific, the sheer amount of data available for model organisms inevitably causes some bias in annotation, or incorrect annotations [184]. In the daffodil data this was seen with the appearance of both the GO category of "singular organisms" and KEGG "pathways of cancer". Further investigation into these terms showed clear links to plant genes such as the *Arabidopsis thaliana* CtBP gene, a member of the D-isomer specific 2 hydroxyacid dehydrogenase family as with Q0VCQ1 in "pathways in cancer" as discussed in chapter 4 [293]. As annotations are often linked to the first gene reported it can lead to mis-annotation as is shown with the Arabidopsis CtBP gene, as although it shares homology to dehydrogenases such as that seen in bovine it is missing the vital amino acids required to dehydrogenase activity[292-294].

Although some incorrect annotations were suggested, the use of GO and KEGG resulted in a suitable background for enrichment studies using DAVID. The DAVID enrichment study resulted in clusters with high enrichment scores for those genes predicted to be involved in secondary metabolite biosynthesis, specifically cytochrome P450s that could be implicated in phenol coupling reactions in the final step of galanthamine biosynthesis.

5.3 Determination of ploidy level of Carlton and Andrew's choice

Both varieties were shown via chromosome counting and the use of pileup_parser.pl to be triploids, this does not agree with the current data on Carlton being tetraploid. However as both varieties used in the project are clonally propagated for the ornamental industry the uneven ploidy level, which is often linked to poor fertility rates would be overlooked in asexual reproduction, as many plants that are bred in this way are triploids [296]. One major example being the cultivated triploid desert bananas of the Cavendish and Gros Michel subgroups [297]. Within narcissus cultivars the original varieties were diploid with only few triploids recorded before 1885, in 1887 the first tetraploid was recorded and now the most popular varieties are triploid and tetraploid [296]. A study by Brandham looked at the high frequency of low fertility triploids within the genus and found that in cultivar divisions 5,6 and 7 the majority were triploids (83.9%, 49% and 72.4%) [298]. Zonneveld also noted that several groups showed both triploid and tetraploids such as *N. poeticus* and *N.tortifolius*[30].

5.4 The effect of polyploidism on gene expression studies

Polyploid events such as genome duplication often result in increased diversity in hybrids allowing them to inhabit novel environments compared to their parents [299]. This "hybrid vigor" results in increased biomass, size, yield and disease resistance [300]. Changes in the new polyploid genomes compared to the parents may allow for evolutionary success via the 'transgressive segregation' of parental alleles (recombination of important alleles in new hybrid) but may also effect gene regulation and therefore expression [299-301].

If multiple genetic or epigenetic changes occur in the master cellular regulators such as promoter or enhancer regions gene expression can be greatly altered [300]. Although gene regulation is dose dependent and therefore these new hybrids have more control over imprinted genes in developing seeds gene expression does not follow the same pattern [299,301]. That is to say that gene expression is not additive, the expression of an allele in a tetraploid for example is not double that seen for the same transcript in a diploid parent [302].

There are several factors that can effect gene expression, transcriptomic shock can occur where there is complete suppression of a transcriptome from one parental genome, tissues specific silencing and up and down regulation can also all effect gene expression patterns in polyploids [299]. The first recognized phenomena of epigenetic changes resulting in expression differences was the silencing of one parental set of rRNA genes while the other parent produced the nucleolus [300]. Normally one parent is considered expression dominant that is to say that on an allelic level it is frequently more dominant [301].

The non-additive expression is linked to regulation, it can be considered to be linked to either trans or cis regulation. If both parental alleles show similar responses to regulatory environmental changes in the hybrid compared to the parental regulatory environment it is considered a trans regulatory effect [300]. A difference in the cis regulatory regions such as promoters and enhancers is often seen with up or down regulation of only one parental allele in the new hybrid [300].

5.5 Prediction of genes involved in galanthamine production

The use of KEGG, GO, EC and BLASTx searches against a GOI database of alkaloid and other key secondary metabolite biosynthetic genes implicated of several homologs in daffodil in the biosynthesis of galanthamine. These are briefly discussed below (see section 4.4.5 for full discussion).

5.5.1 PAL and TYDC

These two enzymes have already been shown to be involved in galanthamine biosynthesis [199,257]. Several transcripts within the daffodil data showed homology to PAL and TYDC, with TYDC being implicated in the enrichment studies as part of "isoquinoline alkaloid biosynthesis". However, none of these transcripts showed SNP or transcript level differences between the two varieties of daffodil and so it is predicted that this step of the pathways does not differ between the two varieties and so is not a point of diversion.

5.5.2 Comp75950_c0_s1 a homolog of C4H in vanillin biosynthesis as a putative transcript involved in the biosynthesis of protocatechuic acid in galanthamine biosynthesis

This transcript showed 22% homology to Q9FQY6_CAP with a fold change of 6.22 (s.d 0.67) in Andrew's Choice. C4H in Capsicum is known to catalyse the reaction of cinnamate to coumarate in the production of vanillin and a similar reaction is seen in galanthamine in the biosynthesis of protocatechuic acid (as discussed in section 4.4) [20]. This is a transcript that needs further investigation to confirm a predicted role in the galanthamine pathway; this and other transcripts for further investigation were discussed in section 5.6.1.

5.5.3 Contig01404 as a homolog for NCS in the biosynthesis of norbelladine

NCS is involved in the production of (S)-norcoclaurine via the addition of 4-HPAA and dopamine in BIA biosynthesis [204,227,268]. The transcript showed homology to C3SBV5_9MAGN in *Thalictrum flavum* (33.7%) and could carry out a similar reaction as that seen for BIA in the conversion of protocatechuic acid and tyramine to norbelladine in galanthamine biosynthesis.

5.5.4 Daff106212 as a homolog for OMTs could be involved in the *O*-methyltransferase catalysed step of norbelladine to 4-*O*-methylnorbelladine

Homology is seen to two members of the COMT subfamily of the OMTs, 6OMT_COPJA in *Coptis japonica* (44%), the (RS)-norcoclaurine synthase implicated in BIA biosynthesis converting (S)-norcoclaurine to (S)-reticuline, and also to Q8GSN1 MOMT_CATRO (52%) from *Catharanthus roseus.* Although OMTs are a large family, the close homology to the enzymes in BIA biosynthesis suggests that this transcript is involved in the conversion of norbelladine to 4-*O*-methylnorbelladine in daffodils. Interestingly, this transcript showed a high fold difference (5.39 ± 0.39) in Andrew's Choice and so could predict a site of diversion or difference in galanthamine biosynthesis between the two varieties linked to their differences in galanthamine levels. However, as discussed in section 5.6.1, predictions like this require experimental confirmations beyond the scope of this work.

5.5.5 HDA57HA0AK3FX and Daff88927 as possible transcripts involved in C-C intermolecular phenol coupling in the final step of galanthamine biosynthesis

HDA57HA0AK3FX shows homology to CYP80 and CYP71A enzymes, both known to be involved in phenol coupling in BIA biosynthesis in *Papaver somniferum, Coptis japonica* and *Eschscholzia californica* [62,217,231]. Daff88927 also shows homology to a CYP71A (48.63%) and a P450 NMT C3SBT0_PAPBR (36%) from *Papaver bracteatum* with a 10-fold increase in Andrew's Choice again suggesting a possible site of diversion in the pathway that requires experimental confirmation.

5.6 Predicted pathway for galanthamine biosynthesis

The analysis of transcriptome data from two varieties of daffodils has led to the prediction of putative transcripts involved in galanthamine biosynthesis and these are shown in a putative pathway in figure 5.1.

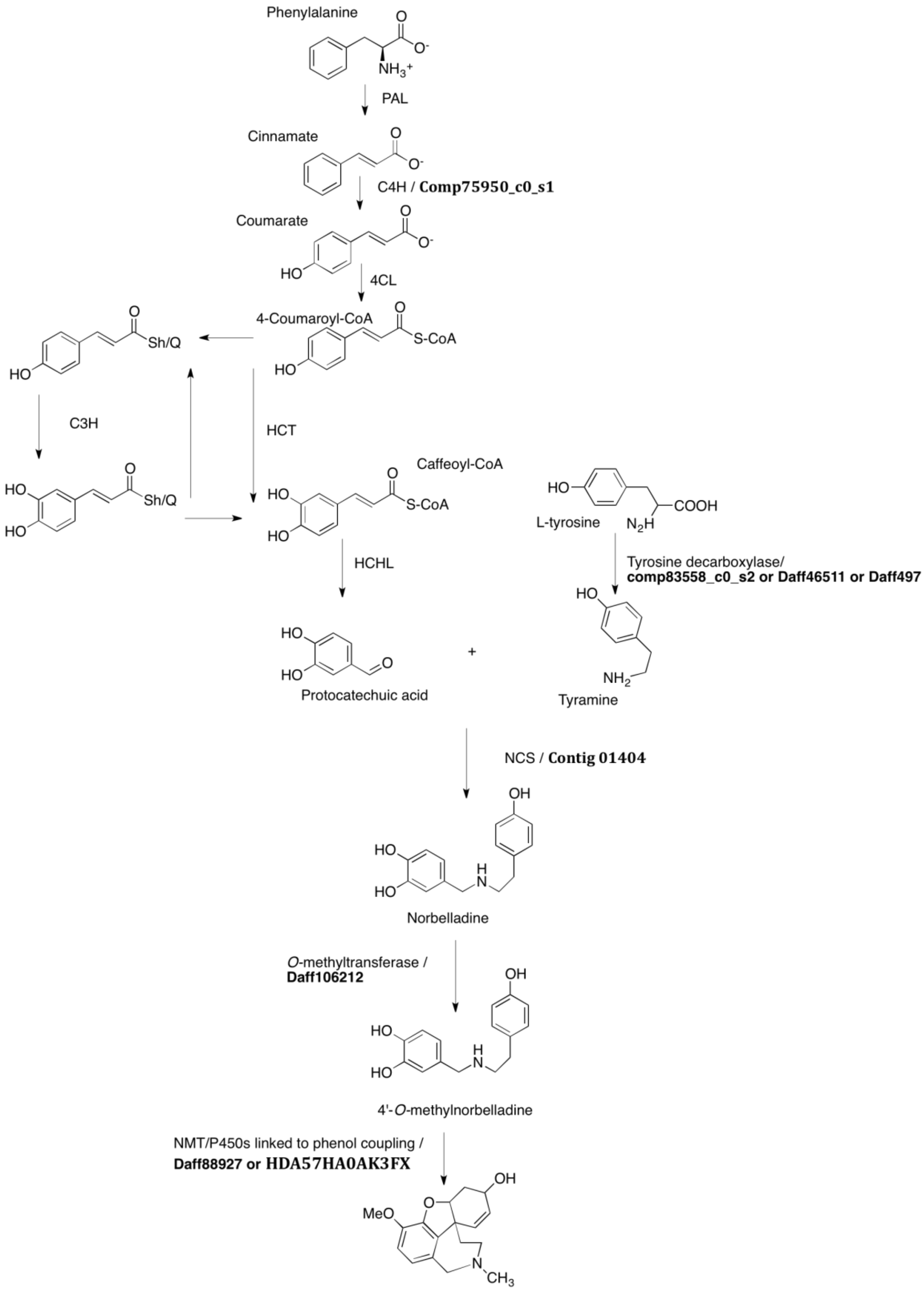


Figure 5-1 Putative pathway for the biosynthesis of galanthamine in daffodils.

The IDs of transcripts predicted in daffodil are shown alongside the predicted enzymes.

5.7 Future Work

The data presented here resulted in the prediction of transcripts that could be involved in alkaloid biosynthesis in daffodils but had several limitations. There is often a compromise between cost and the need to acquire appropriate coverage of the transcriptome, irrespective of the methodologies used [84]. For example, when looking to detect differential expression, inclusion of both biological and technical replicates are considered best practice [303]. Limitations in this project did not allow for repeat sequencing of the samples and although the 454 pyro-sequencing full assembly was made from two different basal plate samples from Carlton, the Illumina sequencing was carried out on only one sample from each variety. However, as can be seen from the results of the transcript level analysis, the predicted fold changes were shown in both sets of sequence data, and in three biological and technical replicates in the qPCR and so the method used can be considered suitable for the scientific question at hand. Comparisons between two cultivars has been used in other studies to predict transcript level differences such as the study by Alagna *et al* on ripening in olives [101].

5.7.1 Expression of candidate genes and subsequent functional analysis on the protein

Enzymes in classes predicted in this project (such as C4H and NMTs) have been confirmed using experimental methods in other plants. Huang *et al* characterized a C4H in *Salvia miltiorrhiza,* through cloning the gene, predicting the protein sequence and identifying conserved domains and structural similarity to previously known C4H genes [254]. Several projects have used both *E. coli* and *S. cerevisiae* to express genes and carry out protein catalysis analysis. Methylenedeoxy bridge-forming CP450s involved in alkaloid production in the Papaveraceae family were analysed using 27 potential substrates [242]. TYDC from *Papaver somniferum* was expressed within *E. coli* and tested for substrate specificity [263]. This could be of particular use with the predicted transcripts in this project as it could confirm their position in the galanthamine pathway. Testing a range of substrates with daffodil enzymes expressed from microbial systems, the chemical products would provide information on which of the

Amaryllidaceae alkaloids were produced. This could, for example, help determine which of HDA57HA0AK3FX and Daff88927 if either is involved in the final phenol-coupling step or if they produce a different alkaloid via one of the other coupling routes.

5.7.2 Microbial engineering to create a galanthamine cell factory

As the levels of alkaloids in daffodils and other medicinally important plants are relatively low, and chemical synthesis is not cost effective, it would be beneficial to reconstruct the galanthamine pathway in a microbial host to meet the demand for the drug [4,26,32]. This strategy has been used for the semi-synthetic commercial production of the anti-malaria drug precursor artemisinin and also for development of a platform for the production of BIAs from simple carbon sources and there is obvious potential to produce other alkaloids such as galanthamine in this way [304,305]. This would first require further characterisation of the enzymes of the pathway way, but could lead to the semi-synthetic production of galanthamine on a commercial scale as is seen with current production of the anticancer terpenoid indole alkaloid vincristine by Eli Lilly [32,306].

5.7.3 Sequencing of genomic DNA to identify regulatory DNA motifs

As sequencing the entire daffodil genome would not be cost effective at the moment, the transcripts predicted in this study could be used as guides to sequence the upstream and downstream regulatory regions. Both computational and experimentally derived sequence motifs in these regions would substantially increase understanding of the regulation of biosynthesis [307]. The interaction of transcription factors with *cis*-acting elements in gene promoter regions is considered an important aspect of regulation and has been studied in another alkaloid producing plant, *Catharanthus roseus.* Two studies isolated and characterized the promoters of the strictosidine synthase and deacetylvindoline-4-O-acetyltransferase genes involved in terpenoid indole alkaloid biosynthesis in this species [308,309]. A similar project would be beneficial to daffodils and could lead to the determination of important regulatory motifs in the alkaloid biosynthesis pathway.

5.7.4 Further investigation into the biological system in daffodil

Transcriptome data from other tissues and under a range of environmental conditions could be used alongside expression data and more extensive alkaloid and other chemical analysis to investigate the compartmentalization of alkaloid production. Galanthamine has been shown to be found in most tissues of daffodils but no information is available of the location of its synthesis[26]. By comparing expression levels in different tissues it may be possible to predict the cellular localization and transport of the alkaloids. This has been extensively studied in poppies. In BIA biosynthesis, gene expression is linked to companion cells and the enzymes are translocated to sieve elements where the alkaloids are synthesised and then stored in laticifers [8].

5.7.5 The development of bioinformatics tools to further evaluate ploidy level in plants

The pileup_parser.pl script in this project predicts ploidy at a nucleotide level. However, it was developed specifically for this project to look for either triploid or tetraploid loci. In order for it to be applicable to other levels of ploidy, several key aspects need to be addressed. Firstly, it does not take into account any sequencing error, quality scores or minimum reads assigned to each locus. Programs used for SNP calling currently use stringent settings to account for these. Developing a similar algorithm in this script would result in more accurate prediction of ploidy level [155,162]. Also, it could be developed alongside experimental SNP frequency data to look at all possible ploidy levels.

5.7.6 The use of SNPs as biomarkers for daffodil breeding linked to alkaloid production

The SNPs found in putative transcripts in galanthamine biosynthesis can be used for future marker development, QTL analysis and genetic linkage. Marker development would be highly beneficial to the commercial extraction of galanthamine as all pharmaceutical grade galanthamine is currently extracted from plants. Daffodils take several years to multiply and so the use of genetic markers linked to alkaloid production could rapidly improve breeding programs for special cultivars. Another plant where this strategy would be beneficial is in cultivars currently being developed for opium poppies to produce the pharmaceutically important precursor thebaine [32].

5.8 Conclusion

In conclusion, this is the first transcriptome comparison study between two varieties of *Narcissus pseudonarcissus* via second-generation sequencing techniques. It provides a platform for the discovery of genes involved in biosynthesis of Amaryllidaceae alkaloids as well as molecular breeding research. The overall aim of this project was to create a reference set of transcripts that could be used to elucidate putative transcripts involved in galanthamine production, to this end the work has resulted in the prediction of several transcripts via numerous methods that should be investigated further to.

However as only one tissue at a specific time point was analysed, albeit one linked to high levels galanthamine production, further investigation is still required to ascertain the localisation of galanthamine production and its pathway. With the rapidly decreasing cost of sequencing and increase in developed techniques for projects of this type if this project was to be carried out again there are several key aspects that could be improved.

Firstly a pooled library from several time points and tissues could be used to give greater insight into whether the pathway is more prevalent in one tissue type or another. Further time points in the life-cycle of the daffodils could also be used to look for changes linked to galanthamine production. Alongside these sample changes it would be beneficial to test the same tissues and time points for galanthamine content so that a co-expression study could be carried out.

Although Carlton and Andrew's choice are related and the comparison of their transcriptomes resulted in the prediction of several gene involved in galanthamine production it would be beneficial to a project of this nature to compare transcript levels to galanthamine levels to look for patterns of co-expression. If a transcript showed both high expression in a specific tissue and time point that was also linked to high levels of galanthamine it would further the evidence for that genes involvement. This is particularly important in a project of this nature as biosynthesis of alkaloids and other secondary metabolites are known to involve from a small number of gene families and have numerous branched pathways.

However, this body of work is an excellent starting point for the elucidation of the enzymes of the galanthamine biosynthetic pathway and shows the validity of similarity searches for the determination of putative genes in non-model plants.

6 Appendix

***PLEASE NOTE:** the majority of the Appendix such as perl scripts and large data tables are supplied on disc. See table 6.1 for description of table and file names on disc.*

6.1 List of perl scripts on appendix disc

All scripts written as part of this project are in a folder labeled "Scripts" on appendix disc all others that were written by other member of Liverpool University are listed but not available on disc. Only the main scripts are shown.

Table 6-1 Perl Scripts.

Script name	Written by
Contigs.pl	J Pulman
Full_annotation.pl	J Pulman
ID_to_FASTA.pl	J Pulman
Joiningfiles.pl	J Pulman
Match_singleton_ID_to_fasta.pl	J Pulman
Pileup_parser.pl	J Pulman
Remove_low_scoring_blast.pl	J Pulman
Removestartinglinesrfam.pl	J Pulman
Singletons.pl	J Pulman
Tabtospace.pl	J Pulman
split_bitseq.pl	J Pulman
split.pl	J Pulman
Annotation_pipeline.pl	Prof. A Hall
Blast_to_hash_search_with_AGI.pl	Prof. A Hall
Extract_nonhuman_reads_fastq.pl	Dr J. Kelly
Unique_agi_ids.pl	Prof. A Hall
coverageStatsSplitByChr_v2.pl	K Ashelford (modified by Laura Gardiner)

6.2 List of data files available on appendix disc

Table 6-2 Data files on appendix disc.

Files can be found in corresponding folders for the section of the thesis denoted in the table.

File name	Relevant section in main thesis
454_Varscan_results	3.4.2.6
Post_TE_Varscan_results	3.4.2.6
454_Andrews_choice_parsing_output	3.4.2.6
454_Carlton_parsing_output	3.4.2.6
Post_TE_Andrews_Choice_parsing_output	3.4.2.6
Post_TE_Carlton_parsing_output	3.4.2.6
454_BitSeq_results	3.4.2.14
Post_TE_BitSeq_results	3.4.2.14
454_GO_annotations	4.3.1.4
Original_GO_annotations	4.3.1.4
Post_TE_removal_GO_annotations	4.3.1.4
454_to_GOI_blast_results	4.3.2
Original_to_GOI_blast_results	4.3.2
Post_TE_removal_to_GOI_blast_results	4.3.2
454_DAVID_EC_GOI_results	4.3.3.3
454_DAVID_Functional_clustering_GOI_results	4.3.3.1
454_DAVID_KEGG_GOI_results	4.3.3.2
Original_DAVID_EC_GOI_results	4.3.3.3
Original_DAVID_KEGG_GOI_results	4.3.3.1
Post_TE_removal_DAVID_EC_GOI_results	4.3.3.2
Post_TE_removal_DAVID_Functional_clustering_GOI_results	4.3.3.3
Post_TE_removal_DAVID_KEGG_GOI_results	4.3.3.1
Data_sheets_for_transcripts_of_interest_from_GOI_search	4.3.3.2

6.3 Galanthamine level data

All data was produced by Dr Xianming Chang formerly of the University of Liverpool and now the Royal Agricultural University.

6.3.1 Introduction

Narcissus pseudonarcissus. Var. Carlton is known to contain levels of galanthamine suitable for extraction for drug production and as such is widely used for this purpose [26]. Andrew's Choice however is a relatively new *N. jonquilla and N. cinel* hybrid that has not been widely studied. It is thought to have much lower levels of galanthamine and was therefore selected as the second variety for this study to compare the transcriptomes of a high producer and a low producer. It is important that the levels of galanthamine produced by this second variety are investigated to support its claim as a low producer. An experiment was designed by Dr X. Chang at the University of Liverpool to determine the alkaloid content of both varieties using a GC-MS method.

6.3.1.1 Experiment 1: Comparison of Andrew's choice and Carlton from three different sites

Eight individuals of each variety were planted at three different sites, one in Newmarket and a high and low altitude plot in Humberside, four individuals of each variety from each plot was harvested on two separate occasions (9 April 2013 and 9 May 2013). 09/04/2013 was at the development stage of goose-necked stage (yellow not open), and 09/05/2013 was at flower dying growth stage (flower not completely dead).

6.3.1.2 Experiment 2: Comparison of the two varieties in the flowering stage of development

Plants grown as described in section 2.3.1 of chapter two. Three plants of each variety were dug up just after flowering and the basal plate sampled in triplicate.

6.3.2 Methodology

The same methodology was carried out for the samples from 6.3.1.1 and 6.3.1.2.

6.3.2.1 Galanthamine extraction for Experiment 1

100mg of frozen tissue was added to 1ml of methanol and adjusted to pH8 with 25% ammonia. The samples are then extracted in an ultrasonic bath for five hours (15minutes on 15 minutes off) and then centrifuged at 10,000rpm for 1 minute. 500ul was removed to a new tube and 10ug of codeine was added as an internal standard.

6.3.2.2 Galanthamine extraction for Experiment 2

100mg of frozen tissue was added to 1ml of methanol and adjusted to pH8 with 25% ammonia. 50ug (10ul) of codeine was added as an internal standard. The samples are then extracted in an ultrasonic bath for two hours (15minutes on 15 minutes off) and then centrifuged at 10,000rpm for 1 minute. 500ul was then transferred to a new tube and 500ul of 2% sulfuric acid added. The neutral compounds were eliminated by duplicate extraction with 500ul chloroform. The organic solvent was evaporated off and the dry extract dissolved in 300ul methanol.

6.3.2.3 Chromatographic techniques

The same methodology was carried out for the samples from 6.3.2.1 and 6.3.2.2. The GC–MS spectra were recorded on a ThermoQuest Trace GC 2000 + PolarisQ MSD operating in EI modeat70eV. ADB-5MS column (30m×0.25mm×0.25μm) was used. The temperature program was: 100–180◦C at 15◦C min−1, 1min hold at 180◦C and 180–300◦C at 5◦C min−1 and 1min hold at 300 ◦ C. The flow rate of carrier gas (Helium) was 0.8 mL min−1. Injector temperature was 280◦C. 1 μL of solutions was injected and a splitless injection was used.

6.3.3 Results

6.3.3.1 Experiment 1

Table 6-3. Comparison of levels of galanthamine for experiment 6.3.2.1 April samples.

Site	**Mean Galanthamine % FW in Carlton (standard deviation/standard error)**	**Mean Galanthamine % FW in Andrew's choice (standard deviation/standard error)**	**T- value**	**P-value (two tailed)**
Newmarket	0.06067(0.021/0.010)	0.01193 (0.004/0.002)	4.54	0.0039
Humberside H	0.08528(0.013/0.007)	0.01343(0.001/0.001)	10.88	$3.6x10^{-5}$
Humberside L	0.06470(0.022/0.011)	0.00659(0.006/0.003)	5.15	0.0021

Table 6-4 Comparison of levels of galanthamine for experiment 6.3.2.1 May samples.

Site	**Mean Galanthamine % FW in Carlton (standard deviation/standard error)**	**Mean Galanthamine % FW in Andrew's choice (standard deviation/standard error)**	**T- value**	**P-value (two tailed)**
Newmarket	0.04860(0.013/0.006)	0.00833(0.004/0.002)	5.916	0.001
Humberside H	0.05049(0.016/0.008)	0.00908(0.003/0.002)	5.06	0.002
Humberside L	0.04503(0.012/0.006)	0.00664(0.003/0.001)	6.39	0.001

6.3.3.2 Experiment 2

Table 6-5 Comparison of levels of galanthamine from experiment 6.3.2.2.

Mean Galanthamine % FW in Carlton (standard deviation/standard error)	**Mean Galanthamine % FW in Andrew's choice (standard deviation/standard error)**	**T- value**	**P-value (two tailed)**
0.01300 (0.004/0.002)	0.03983(0.008/0.003)	6.48	0.0002

6.3.4 Discussion

Both experiments showed clear differences between the two varieties of daffodil suggesting that Carlton produces significantly higher levels of galanthamine. As the method of extraction was different the two experiments could not be directly compared for galanthamine levels but do both show a clear significant difference between the two varieties. Therefore the two cultivars chosen are suitable candidates as the aim of this project is to compare the transcriptome of two cultivars with significant differences in galanthamine level to look for putative transcripts involved in its biosynthesis.

6.4 Chromosome counting for Carlton and Andrew's choice varieties of daffodil

6.4.1 Introduction

As the ploidy level of *Narcissus pseudonarcissus* varieties can range from diploid to nonoploid is important that the ploidy level of the varieties used in this project were known [30]. Unfortunately neither variety was included in Zonneveld's 2008 study and so the amount of DNA is unknown. An estimation of the ploidy level can be determined via chromosome counting. Although Carlton has been investigated before and shown to be tetraploid (2n=28), the number of cultivars in daffodils and the lack of pedigree information available for Carlton it was decided that the ploidy level of the particular individuals used in this study should be confirmed [31].

6.4.2 Method

Bulbs of both varieties were grown as described in section 2.3.1.

Root removal, preparation and chromosome counting was carried out under the supervision and instruction of Dr Hugh McAllister from Ness Gardens. The method he produced was based on numerous publications and is as follows.

6.4.2.1 Reagents

Supplied by Dr Hugh McAllister, University of Liverpool.

Pretreatment solution: Saturated aqueous solution of 1-bromo-naphthalein (supernatant over small quantity of reagent - keep topped up with tap water).

Fixative: Freshly mixed 3 : 1, 95% ethanol (industrial): glacial acetic acid.

Hydrolyzing solution: 1M hydrochloric acid.

Storage medium: 70% ethanol.

Mountant: 2 : 1 lactic acid : propionic acid.

6.4.2.2 Pretreatment

Rapidly growing roots of each variety (cleaned in water to remove soil) were placed in separate screw top glass vials with pretreatment solution (roots covered) and left in the cold room (4°C) overnight.

6.4.2.3 Fixation

The roots were removed from the pretreatment solution and added to vials containing freshly mixed fixative and stored for overnight in cold room (4°C). Root tips can be stored at this stage for many months if not years.

6.4.2.4 Hydrolysis

Root tips were removed from fixative and place in screw top plastic 1.5ml eppendorfs of 1M HCl at 60°C on a heat block for 5 min, then transferred to 70% ethanol for storage.

6.4.2.5 Slide preparation

A dissecting microscope was used and the cytoplasmic region of the root tip was removed with needles and then the remaining section cut in to small pieces on a glass slide. A drop of mountant was added and the sample tapped out to a monolayer under a cover slip and then squashed with medium force (glass cover slip shouldn't break but enough force is needed to flatten out cells but not burst them). The resulting slide was then examined by phase contrast microscopy and any whole cells with chromosomes in the appropriate phase (dividing) for a viable count were counted by three different individuals.

6.4.3 Results and discussion

6.4.3.1 Carlton:

Three slides produced cells that were suitable for counting, although the counting could not be performed to a degree of + or – 1 the chromosome number was clearly over 14 and less than 28. All three members of the team counted each slide several time, the number ranged from 18-22 and so Carlton is predicted to be triploid. It was not possible to get a suitable photo showing the individual chromosomes.

6.4.3.2 Andrew's choice:

Only two slides produced cells that were suitable for counting, it is clear in figure 6.1 that the cells used were whole and that the chromosomes were very elongated. This elongation and large size of daffodil chromosomes makes carrying out a count to a degree of + or – 1 very difficult. As with the Carlton slides the count was estimated to be well above 14 and below 28 again suggesting a triploid. (3n=21).

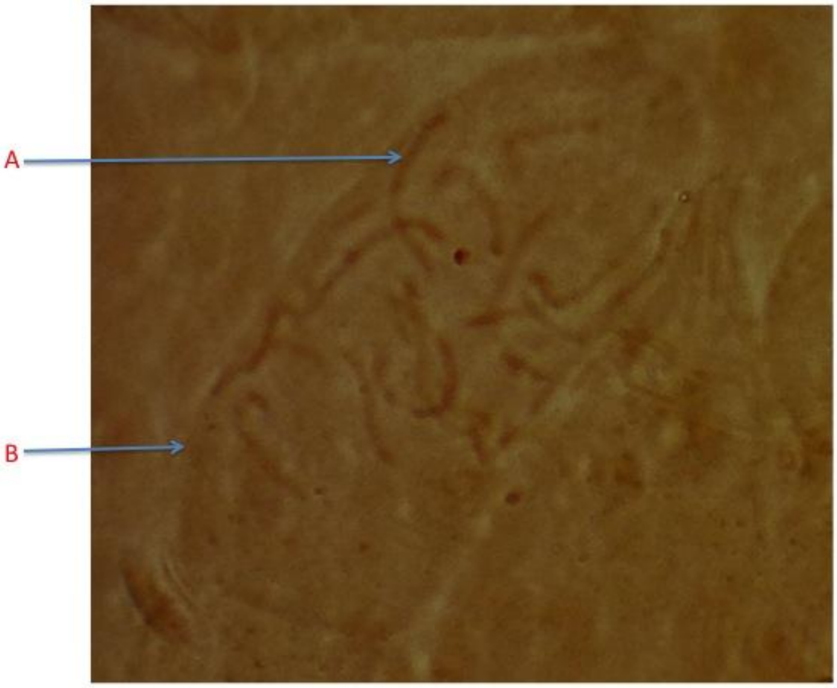


Figure 6-1 Phase contract microscope image of Andrew's Choice cells.

The cells show individual chromosomes that can be counted to estimate ploidy. A is a chromosome and B shows the outline of the cell wall.

6.5 Andrew's Choice RNA extraction trial results and discussion

The results of the trial can be seen in table 6.6. It is clear that the best method of extraction is using the Analytik-Jena InnuPREP plant RNA Kit with the PL lysis solution. The kit gives high consistent yields and consistent 260/280 ratios. Therefore this method was used alongside a DNase step (Qiagen DNase kit) to extract the RNA to be used to create a cDNA library for the Andrew's Choice variety. The extraction resulted in two samples with yields of 696 ng/µl and 740 ng/µl. Both samples had 260/280 ratios within the range (1.8-2.3) suggesting pure RNA and 260/230 ratios suggesting no protein contamination. The samples were frozen and stored at -80°C until needed for cDNA library preparation.

Table 6-6 RNA extraction trial results.

Date	**Experiment**	**Sample**	**Method**	**ng ml^{-1}**	**260/280**	**260/230**	**Gel results**	**Comments**
07/12/2011	1	Andrew's Choice, Liverpool, April 2011	CTAB followed by RNeasy clean up	920	2.2	2.3	Ran on Qubit ® 2.0 fluorometer looked intact and concentration 700ngul^{-1}	Carried forward to mRNA selection using Dynabeads method. Failed to show any rRNA depletion and yield too low to continue
07/12/2011	2	Andrew's Choice, Liverpool, April 2011	CTAB followed by RNeasy clean up	564	2.2	1.6	ND	260/230 value below 2, which suggests impure RNA. Not used for further work
07/02/2012	3	Andrew's Choice, Liverpool, April 2011	CTAB followed by RNeasy clean up	2116	2.2	2.3	Mostly intact some degradation	This was a practise run to show Xianming the methodology. Good yield but shows that CTAB method is not consistent.
07/02/2012	4	Andrew's Choice, Liverpool, April 2011	CTAB followed by RNeasy clean up	2750	2.2	2	Mostly intact some degradation	This was a practise run to show Xianming the methodology. Good yield but shows that CTAB method is not consistent.
14/02/2012	5	Harvest 3 high field replicate 2 Andrew's choice	MoBio Power Plant RNA extraction kit followed by RNeasy clean up with optional DNase on column step	23	2	0.5	Yield too low to visualise.	260/230 value suggested very poor degraded RNA with numerous contaminants. Could be caused by DNase step. Future plan to test Mo Bio kit without any DNase or clean up.
14/02/2012	6	Harvest 3 high field replicate 2 Andrew's choice	MoBio Power Plant RNA extraction kit followed by RNeasy clean up with optional DNase on column step	64	2.1	0.5	Yield too low to visualise	260/230 value suggested very poor degraded RNA with numerous contaminants. Could be caused by DNase step. Future plan to test Mo Bio kit without any DNase or clean up.
17/02/2012	7	Harvest 3 high field replicate 2 Andrew's choice	MoBio Power Plant RNA extraction kit	120	2	2.2	ND	Graphs and 260/230 suggest quality RNA but low yields. Also DNA contamination is still present (221.72 ng ml^{-1})

17/02/2012	8	Harvest 3 high field replicate 2 Andrew's choice	MoBio Power Plant RNA extraction kit	124	1.9	1.5	ND	More contamination poorer quality than sample 7. Also DNA contamination (162.96ngul^{-1})
17/02/2012	9	RNA extracted in experiment 7	MoBio RTS DNase	141	1.8	1.6	Highly degraded and DNA contamination	Contaminated and still DNA Present (178.41ngul^{-1})
17/02/2012	10	RNA extracted in experiment 8	MoBio RTS DNase	120	1.7	1.3	Highly degraded and DNA contamination	Contaminated and still DNA Present (127.60ngul^{-1})
17/02/2012	11	RNA extracted in experiment 7	RNeasy clean up	141	2	2.1	Better quality than experiment 9 and 10 but still degraded	DNA present (176.57ngul-1)
17/02/2012	12	RNA extracted in experiment	RNeasy clean up	110	1.4	2	Better quality than experiment 9 and 10 but still degraded	DNA present (137.43ngul-1)
28/02/2012	13	Harvest 3 high field replicate 3 AC	CTAB followed by RNeasy clean up	553	2.2	2.4	Reasonable quality. Too low a yield	Continues to show inconsistent results.
28/02/2012	14	Harvest 3 high field replicate 3 AC	CTAB followed by RNeasy clean up	99	2.1	1	Too low to show up poor quality	260/230 suggests poor quality and very low yield.
29/02/2012	15	Harvest 3 low field replicate 2 AC	InnuSPEED Plant RNA kit	221	1.8	1.3	Too low a yield to show up	260/230 too low suggests contaminants
29/02/2012	16	Harvest 3 low field replicate 2 AC	InnuSPEED Plant RNA kit	182	1.8	1.3	Too low a yield to visualise	260/230 too low. However experiment 15 and 16 are more consistent yields than other experiments.
29/02/2012	17	Harvest 3 low field replicate 2 AC	InnuPREP Plant RNA kit	864	2.2	2.3	Higher yields, consistent, good quality.	Shows promise. Good quality RNA on gel.
29/02/2012	18	Harvest 3 low field replicate 2 AC	InnuPREP Plant RNA kit	1452	2.2	2.3	Higher yields, consistent, good quality.	Quality good.

6.6 List of Plants and Genes in GOI database

Table 6-7 Plants and Genes included in GOI database.

UNIPROT ID	Enzyme	Species
Q9LVY1	Tyrosine aminotransferase	Arabidopsis thaliana
Q9FN30	Probable aminotransferase TAT2	Arabidopsis thaliana
Q9SK47	Probable aminotransferase TAT3	Arabidopsis thaliana
D3K4J1	Tyrosine aminotransferase	Papaver somniferum
G8HAA8	PLP-dependent aminotransferase	Papaver somniferum
G8HAB3	PLP-dependent aminotransferase	Papaver somniferum
G8HAB0	PLP-dependent aminotransferase	Papaver somniferum
G8HAB2	PLP-dependent aminotransferase	Papaver somniferum
G8HAA9	PLP-dependent aminotransferase	Papaver somniferum
G8HAB1	PLP-dependent aminotransferase	Papaver somniferum
Q8GUE9	Tyrosine aminotransferase	Solenostemon scutellarioides
Q8RY79	Tyrosine decarboxylase 1	Arabidopsis thaliana
P54770	Tyrosine/DOPA decarboxylase 3	Papaver somniferum
P54768	Tyrosine/DOPA decarboxylase 1	Papaver somniferum
P54769	Tyrosine/DOPA decarboxylase 2	Papaver somniferum
P54771	Tyrosine/DOPA decarboxylase 5	Papaver somniferum
B6E2Z2	Norcoclaurine synthase	Papaver somniferum
C3SBV6	Norcoclaurine synthase	Thalictrum flavum
C3SBV5	Norcoclaurine synthase (Fragment)	Thalictrum flavum
A2A1A0	S-norcoclaurine synthase 1	Coptis japonica
A2A1A1	S-norcoclaurine synthase 2	Coptis japonica
C3SBU1	Putative norcoclaurine synthase	Papaver bracteatum
C3SBT7	Putative norcoclaurine synthase	Papaver bracteatum
Q4QTJ1	S-norcoclaurine synthase 2	Papaver somniferum
Q4QTJ2	S-norcoclaurine synthase 1	Papaver somniferum
D2SMN3	S-norcoclaurine synthase 2	Argemone mexicana
C3SBS5	Pathogenesis-related (PR)-10-related norcoclaurine synthase-like protein	Eschscholzia californica
C3SBS4	Pathogenesis-related (PR)-10-related norcoclaurine synthase-like protein	Eschscholzia californica
D2SMN1	S-norcoclaurine synthase 1	Argemone mexicana
Q39224	Protein SRG1	Arabidopsis thaliana
Q42393	Sn-1 protein	Capsicum annuum
Q9LEL6	(RS)-norcoclaurine 6-O-methyltransferase	Coptis japonica
Q9FK25	Flavone 3'-O-methyltransferase 1	Arabidopsis thaliana
Q9FQY8	Caffeic acid 3-O-methyltransferase	Capsicum annuum
O81646	Caffeic acid 3-O-methyltransferase	Capsicum chinense
B5LAT9	Putative caffeoyl-CoA 3-O-methyltransferase	Capsicum annuum
Q6WUC0	Catechol O-methyltransferase	Papaver somniferum
I3PLQ7	O-methyltransferase	Papaver somniferum
I3PLQ5	O-methyltransferase 1	Papaver somniferum
Q6WUC1	(R,S)-norcoclaurine 6-O-methyltransferase	Papaver somniferum
Q6WUC2	(R,S)-reticuline 7-O-methyltransferase	Papaver somniferum
C7SDN9	Norreticuline-7-O-methyltransferase	Papaver somniferum
C3SBT1	Putative norcoclaurine 6-O-methyltransferase (Fragment)	Papaver bracteatum
C3SBT9	Putative norcoclaurine 6-O-methyltransferase	Papaver bracteatum
I3PLQ6	O-methyltransferase 2	Papaver somniferum
I3V6A7	Scoulerine-9-O-methyltransferase	Papaver somniferum
Q7XB09	S-adenosyl-L-methionine:norcoclaurine 6-O-methyltransferase	Papaver somniferum
Q7XB11	S-adenosyl-L-methionine:3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase 1	Papaver somniferum
Q7XB10	S-adenosyl-L-methionine:3'-hydroxy-N-methylcoclaurine 4'-O-methyltransferase 2	Papaver somniferum
C3SBT3	Reticuline 7-O-methyltransferase-like protein	Papaver bracteatum
I3V6A8	Narcotoline-O-methyltransferase	Papaver somniferum
C3SBW1	Reticuline 7-O-methyltransferase-like protein	Papaver bracteatum

Q7XB08	S-adenosyl-L-methionine:coclaurine N-methyltransferase	Papaver somniferum
Q8L9U0	Coclaurine N-methyltransferase	Arabidopsis thaliana
Q5C9L6	(S)-coclaurine N-methyltransferase	Thalictrum flavum subsp. glaucum
Q948P7	Coclaurine N-methyltransferase	Coptis japonica
C3SBU8	Coclaurine N-methyltransferase (Fragment)	Thalictrum flavum
C3SBV3	Coclaurine N-methyltransferase-like protein (Fragment)	Thalictrum flavum
O64899	(S)-N-methylcoclaurine 3'-hydroxylase isozyme 1 (Fragment)	Eschscholzia californica
O64900	(S)-N-methylcoclaurine 3'-hydroxylase isozyme 2	Eschscholzia californica
Q9FXW4	Probable (S)-N-methylcoclaurine 3'-hydroxylase isozyme 2	Coptis japonica
O64901	(S)-N-methylcoclaurine 3'-hydroxylase	Eschscholzia californica
B9VRK4	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver orientale
B9VRK5	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver rhoeas
C3SBS0	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Eschscholzia californica
Q9M7I3	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver somniferum
B9VRK1	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver somniferum
Q9SP06	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver somniferum
B9VRK3	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver nudicaule
B9VRK2	(S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver bracteatum
C3SBT0	Putative (S)-N-methylcoclaurine 3'-hydroxylase (Fragment)	Papaver bracteatum
Q5C9L5	(S)-N-methylcoclaurine 3'-hydroxylase	Thalictrum flavum subsp. glaucum
P80969	Tyramine N-feruloyltransferase 10/30	Nicotiana tabacum
Q9SMB8	Tyramine N-feruloyltransferase 4/11	Nicotiana tabacum
Q9ATJ3	Hydroxycinnamoyl-CoA: serotonin N-(Hydroxycinnamoyl)transferase	Capsicum annuum
Q9ZV06	At2g39020	Arabidopsis thaliana
P35510	Phenylalanine ammonia-lyase 1	Arabidopsis thaliana
P45725	Phenylalanine ammonia-lyase 3	Arabidopsis thaliana
Q9SS45	Phenylalanine ammonia-lyase 4	Arabidopsis thaliana
Q65CJ7	Hydroxyphenylpyruvate reductase (HPPR)	Solenostemon scutellarioides
B1GV49	Cinnamate-4-hydroxylase	Arabidopsis thaliana
B1GV39	Cinnamate-4-hydroxylase	Arabidopsis thaliana
B1GV36	Cinnamate-4-hydroxylase	Arabidopsis thaliana
B1GV37	Cinnamate-4-hydroxylase	Arabidopsis thaliana
B1GV38	Cinnamate-4-hydroxylase	Arabidopsis thaliana
B1GV55	Cinnamate-4-hydroxylase (Fragment)	Arabidopsis thaliana
B1GV56	Cinnamate-4-hydroxylase (Fragment)	Arabidopsis lyrata
Q9FQY6	Cinnamic acid 4-hydroxylase	Capsicum annuum
Q9S725	4-coumarate--CoA ligase 2	Arabidopsis thaliana
Q9LU36	4-coumarate--CoA ligase 4	Arabidopsis thaliana
Q42524	4-coumarate--CoA ligase 1	Arabidopsis thaliana
Q9S777	4-coumarate--CoA ligase 3	Arabidopsis thaliana
B1GUZ3	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GUZ2	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GUV7	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GUV0	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GUW8	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GV02	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GUZ8	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GUV5	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GUW5	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GUZ6	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GUW0	4-cumarate-COA-ligase (Fragment)	Arabidopsis thaliana
B1GUZ7	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
B1GV07	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana

B1GV10	Cinnamyl alcohol dehydrogenase	Arabidopsis thaliana
A8MS69	4-coumarate--CoA ligase 1	Arabidopsis thaliana
F4I9T8	4-coumarate--CoA ligase 3	Arabidopsis thaliana
Q9FQY7	4-coumarate:coenzyme A ligase	Capsicum annuum
P13114	Chalcone synthase	Arabidopsis thaliana
Q460R0	Chalcone synthase family protein	Arabidopsis thaliana
Q705N7	Chalcone synthase (Fragment)	Arabidopsis thaliana
Q64HU3	Chalcone synthase (Fragment)	Arabidopsis thaliana
Q4JNW5	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNW9	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q705N9	Chalcone synthase (Fragment)	Arabidopsis thaliana
Q460R2	Chalcone synthase family protein	Arabidopsis thaliana
Q460R8	Chalcone synthase family protein	Arabidopsis thaliana
Q460Q4	Chalcone synthase family protein	Arabidopsis thaliana
Q460S8	Chalcone synthase family protein	Arabidopsis thaliana
Q460S2	Chalcone synthase family protein	Arabidopsis thaliana
Q460R3	Chalcone synthase family protein	Arabidopsis thaliana
Q460T2	Chalcone synthase family protein	Arabidopsis thaliana
Q460Q3	Chalcone synthase family protein	Arabidopsis thaliana
Q5NDK6	Chalcone synthase	Arabidopsis croatica
Q5FBU9	Mutant protein of Chalcone synthase	Arabidopsis thaliana
Q5FBU5	Mutant protein of Chalcone synthase	Arabidopsis thaliana
Q5FBU6	Mutant protein of Chalcone synthase	Arabidopsis thaliana
Q460V1	Chalcone synthase family protein	Arabidopsis thaliana
Q460V0	Chalcone synthase family protein	Arabidopsis thaliana
Q4JNU5	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNU1	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q460T8	Chalcone synthase family protein	Arabidopsis thaliana
Q460V5	Chalcone synthase family protein	Arabidopsis thaliana
Q4JNW8	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNT9	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNU6	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNU2	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNW2	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNW6	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q460U2	Chalcone synthase family protein	Arabidopsis thaliana
Q460W0	Chalcone synthase family protein	Arabidopsis thaliana
Q4JNW4	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
Q4JNV9	Chalcone synthase family protein (Fragment)	Arabidopsis thaliana
I3WWC8	Chalcone synthase	Capsicum annuum
C0LF60	Chalcone synthase (Fragment)	Capsicum annuum
A5A369	Chalcone synthase	Solenostemon scutellarioides
C9VWQ6	CHS (Fragment)	Solenostemon scutellarioides
Q9FQY8	Caffeic acid 3-O-methyltransferase	Capsicum annuum
O81646	Caffeic acid 3-O-methyltransferase	Capsicum chinense
Q9FK25	Flavone 3'-O-methyltransferase 1	Arabidopsis thaliana
O49499	Caffeoyl-CoA O-methyltransferase 1	Arabidopsis thaliana
Q6WUC0	Catechol O-methyltransferase	Papaver somniferum
Q6WUC1	(R,S)-norcoclaurine 6-O-methyltransferase	Papaver somniferum
C7SDN9	Norreticuline-7-O-methyltransferase	Papaver somniferum
Q6WUC2	(R,S)-reticuline 7-O-methyltransferase	Papaver somniferum
I3PLQ7	O-methyltransferase	Papaver somniferum
G3FDY1	Caffeic acid O-methyltransferase	Salvia miltiorrhiza
G3FDY0	Caffeic acid O-methyltransferase	Salvia miltiorrhiza

6.7 Primers for RT-PCR, Sanger sequencing and qPCR

Table 6-8 Primers for RT-PCR, Sanger sequencing and qPCR.

Contig id	Forward Primer	TM	Reverse Primer	TM	Product length (bp)	Original or nested	Sent for Sanger sequencing	Used in qPCR
Contig01152	CGAGTCTATAAAACCATTCATCGAG	59.7	TTCAATTATCACTTTGGCAACTACA	56.5	223	original	Yes	No
HDA57HA01BEL8O	CTGATCTCCTTTATAACAGCTTCCA	59.7	ATCGGCATTATAAATCTGAAAACAG	56.4	310	Original	No	no
	TTCCAGGTAGCTCATTTTGC	55.3	GATTTTTGTGGAAGCTCTGC	55.3	264	nested	No	No
HDA57HA010VJD	TTGCACCAATTTGGAACTCA	53.2	GTGCTACGGAAACACCCATT	57.3	103	Original	yes	no
Daff74484	AGCTACGTAGTCAATCTCATTGGTC	60	ATGTCTGTAATTCCATTCGTAGAGC	59	402	Original	Yes	yes
Daff88927	CAGTTGGTTTAATTCATCTCTGCTT	58.1	ATGACAGAATTCTAGCAGCTTTGTT	58.1	332	Original	Yes	yes
Daff106212	ATTCCAGCTAAAGAAAAAGGAGGTA	58.1	TCTTGAACTCATTCTCAGTTCTCTG	58.1	163	original	Yes	yes
COMP75950_C0_S1	GATTGTGGCAGACCCAGAGT	59.4	CGGAGATCCAGACGTAGGAG	61.4	97	original	No	No
COMP97312_C0_S1	ATCGACCAACGAAGCTAGGA	57.3	AATTTGCTTCCCCCTCACAT	55.3	145	original	Yes	No
COMP99544_C0_S1	TGCTCTCCTCCAACAGTTCA	57.3	CCGAGGATTACTTGCCATTC	57.3	126	original	Yes	No
COMP100406_C1_S2	AATTGAGCGCAATCCAAGAG	55.3	TTCACACAACTGGGAAGCAG	57.3	107	original	Yes	No
COMP100760_C0_S2	CAAGCTTCCACCTTCTCCTC	59.4	ATGAGGGGGCCATACTTCTT	57.3	117	original	Yes	No
Comp101410c0s3	AGATCGGGGACTTCGAAAAT	55.3	GGCGACGAGCTAGATACGAG	61.4	92	original	No	Yes
Contig01404	GAGATGGAGCACTTGGAAGC	59.4	CCACgAAATCAAGACGTTCC	57.3	87	original	No	Yes
HDA57HA01B3O58	CTTTGCTGTGACTGGGATGA	57.3	ATCACCTTTCCGTTGTCAGG	57.3	83	original	No	Yes
HDA57HA01AK3FX	ACCATGCCAAGAAATTGGAG	57.3	TCTGTCATGCTCCGTTCAAG	55.3	190	original	No	Yes
Contig01885	TTTGGAGATCGACCTTGTCC	59.4	CGAAACTCGCATGGACTACA	57.3	83	original	No	Yes
Contig02587	CCAGATTCaGGTGTTGAGCA	57.3	GGAGAATCGCTCGCTAGATG	59.4	82	original	No	Yes
Contig03502	TGCCATCAAGTTTGTTCGAG	55.3	TTCCTgAGCTCGTGaAAgGT	57.3	93	original	No	Yes
HDA57HA01AW38A	TTACACCTCCACGAGCAACA	57.3	CCTCACACCTACCACAGCAA	59.4	95	original	no	Yes
ACTIN	GATAGAACCTCCAATCCAAACACTA	59.7	GTGTGATGTGGATATTAGGAAGGAC	61.3	184	original	yes	yes

6.8 qPCR Plate set up

One transcript of interest was run on each plate alongside actin for comparison.

Table 6-9 qPCR plate layout.

C1+ve	**C1+ve**	**C1+ve**	**C1-ve**	**C1-ve**	**C1-ve**	**C1+ve**	**C1+ve**	**C1+ve**	**C1-ve**	**C1-ve**	**C1-ve**
C2+ve	**C2+ve**	**C2+ve**	**C2-ve**	**C2-ve**	**C2-ve**	**C2+ve**	**C2+ve**	**C2+ve**	**C2-ve**	**C2-ve**	**C2-ve**
C3+ve	**C3+ve**	**C3+ve**	**C3-ve**	**C3-ve**	**C3-ve**	**C3+ve**	**C3+ve**	**C3+ve**	**C3-ve**	**C3-ve**	**C3-ve**
AC1+ve	**AC1+ve**	**AC1+ve**	**AC1-ve**	**AC1-ve**	**AC1-ve**	**AC1+ve**	**AC1+ve**	**AC1+ve**	**AC1-ve**	**AC1-ve**	**AC1-ve**
AC2+ve	**AC2+ve**	**AC2+ve**	**AC2-ve**	**AC2-ve**	**AC2-ve**	**C2+ve**	**AC2+ve**	**AC2+ve**	**AC2-ve**	**AC2-ve**	**AC2-ve**
AC3+ve	**AC3+ve**	**AC3+ve**	**AC3-ve**	**AC3-ve**	**AC3-ve**	**C3+ve**	**AC3+ve**	**AC3+ve**	**AC3-ve**	**AC3-ve**	**AC3-ve**
H_2O	**H_2O**	**H_2O**				**H_2O**	**H_2O**	**H_2O**			

C = Carlton cDNA
1, 2 or 3= biological replicate
+ve= cDNA
-ve= reaction mix from cDNA synthesis without Reverse Transcriptase
H_2O= RO water instead of cDNA in PCR reaction

▨ = PCR ran with actin primers   □ = PCR ran with transcript specific primers of interest primers

6.9 ClustalO Alignment from UniProt Alignment program

```
CLUSTAL O(1.2.1) multiple sequence alignment

SP|P45724|PAL2_ARATH MDQIEAMLC----GGGEKTKVAVTTKTLADPLNWGLAADQMKGSHLDEVKKMVEEYRRPV 56
SP|P45727|PALY_PERAE ------------------------------------------------------------
SP|P45729|PAL3_PETCR MAYVNGTTNGH--ANGNGL---DLCMKKEDPLNWGVAAEALTGSHLDEVKRMVAEYRKPV 55
SP|P45726|PALY_CAMSI MDSTTAIGNGV--GSGGSP---GFCL--KDPLNWGVAAEAMKGSHLEEVKGMVEEFRKPV 53
SP|Q42609|PALY_BROFI ---------------MEVSKENGLCLQGRDPLNWGAAAAELQGSHLDEVKKMVEEFRRPV 45
SP|O64963|PAL1_PRUAV MATNSIKQNGHKNGSVELP---ELCIK-KDPLNWGVAAETLKGSHLDEVKRMVAEYRKPV 56


SP|P45724|PAL2_ARATH VNLGGETLTIGQVAAISTVG-GSVKVELAETSRAGVKASSDWVMESMNKGTDSYGVTTGF 115
SP|P45727|PALY_PERAE ------------------------------------------MESMDKGTDSYGVTTGF 17
SP|P45729|PAL3_PETCR VKLEGETLTISQVAAISARDDSGVKVELSEEARAGVKASSDWVMDSMNKGTDSYGVTTGF 115
SP|P45726|PALY_CAMSI VRLGGETLTISQVAAIAVRG-SEVAVELSESAREGVKASSDWVMESMNKGTDSYGVTTGF 112
SP|Q42609|PALY_BROFI VKLEGVKLKISQVAAVAFGG-GASAVELAESARAGVKASSDWVLESVDKGTDSYGVTTGF 104
SP|O64963|PAL1_PRUAV VKLGGESLTISQVAAIATH-DSGVKVELSESARAGVKASSDWVMDSMSKGTDSYGVTTGF 115
                                          ::*:.************

SP|P45724|PAL2_ARATH GATSHRRTKNGTALQTELIRFLNAGIFGNTKET-CHTLPQSATRAAMLVRVNTLLQGYSG 174
SP|P45727|PALY_PERAE GATSHRRTKQGGALHKELIRFLNAGIFGTNGESG-HTLAPSATRAAMLVRINTLLQGYSG 76
SP|P45729|PAL3_PETCR GATSHRRTKQGGALQKELIRFLNAGIFGSGAEAGNNTLPHSATRAAMLVRINTLLQGYSG 175
SP|P45726|PALY_CAMSI GATSHRRTKEGGALQKELIRFLNAGIFGNGTES-CHTLPQSATRAAMLVRINTLLQGYSG 171
SP|Q42609|PALY_BROFI GATSHRRTKQGGALQKELIKFLNAGIFGSGN---SNTLPSAATRAAMLVRINTLLQGYSG 161
SP|O64963|PAL1_PRUAV GATSHRRTKQGAALQKELIRFLNAGVFGSTKESG-HTLPHQATRAAMLVRINTLLQGYSG 174
              *********.* **..***.*****.**.     .**   *********.*********

SP|P45724|PAL2_ARATH IRFEILEAITSLLNHNISPSLPLRGTITASGDLVPLSYIAGLLTGRPNSKATGPDGESLT 234
SP|P45727|PALY_PERAE IRFEILEAITSLLNHSITPCLPLRGTITASGDLVPLSYIAGMLTGRPNSKGDWPDGKEID 136
SP|P45729|PAL3_PETCR IRFEILEAITKFLNHNITPCLPLRGTITASGDLVPLSYIAGLLTGRPNSKAVGPTGVTLS 235
SP|P45726|PALY_CAMSI IRFEILEAISKFLNNNITPCLPLRGTITASGDLVPLSYIAGLLTGRHNSKAVGPTGEILH 231
SP|Q42609|PALY_BROFI IRFEILKAIATLLNKNITPCLPLRGTITASGDLVPLSYLAGILTGRPNSKARTPNGSTVD 221
SP|O64963|PAL1_PRUAV IRFEILEVITKFLNNNVTPCLPLRGTITASGDLVPLSYIAGMLTGRPNSKAVGPDGQTLS 234
              ******.:*.:.**..:*.*****************.**.**** ***.  * *  :

SP|P45724|PAL2_ARATH AKEAFEKAGISTGFFDLQPKEGLALVNGTAVGSGMASMVLFEANVQAVLAEVLSAIFAEV 294
SP|P45727|PALY_PERAE AGEAFRLAGIPSGFFELQPKEGLALVNGTAVGSGLASMVLFEANVLSVLSEVISAIFCEV 196
SP|P45729|PAL3_PETCR PEEAFKLAGVEGGFFELQPKEGLALVNGTAVGSGMASMVLFEANILAVLAEVMSAIFAEV 295
SP|P45726|PALY_CAMSI PKEAFRLAGVEGGFFELQPKEGLALVNGTAVGSGLASMVLFEANILAVLSEVLSAIFAEV 291
SP|Q42609|PALY_BROFI ATTAFRLAGISSGFFDLQPKEGLALVNGTAVGSGVASIVLFETNILAVMAELLSALFCEV 281
SP|O64963|PAL1_PRUAV AAEAFEFVGINSGFFELQPKEGLALVNGTAVGSGLASTVLFDTNILALLSEILSAIFAEV 294
                **..*:  ***.*****************.** ***::*: ::::*::**:*.**

SP|P45724|PAL2_ARATH MSGKPEFTDHLTHRLKHHPGQIEAAAIMEHILDGSSYMKLAQKVHEMDPLQKPKQDRYA- 353
SP|P45727|PALY_PERAE MQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYMKVAKKLHELDPLQKPKQDPYAA 256
SP|P45729|PAL3_PETCR MQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSAYVKAAQKLHEMDPLQKPKQDRYA- 354
SP|P45726|PALY_CAMSI MQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYVKAAQKLHEMDPLQKPKQDRYA- 350
SP|Q42609|PALY_BROFI MQGKPEFTDHLTHKLKHHPGQIEAAAVMEHILEGSSYMKMAKKLHEMDPLQKPKQDRYA- 340
SP|O64963|PAL1_PRUAV MQGKPEFTDHLTHKLKHHPGQIEAAAIMEHILDGSSYVKAAKKLHEQDPLQKPKQDRYA- 353
              *.***********.***********.*****.**.*.* *.*.** ********* **

SP|P45724|PAL2_ARATH LRTSPQWLGPQIEVIRQATKSIEREINSVNDNPLIDVSRNKAIHGGNFQGTPIGVSMDNT 413
SP|P45727|PALY_PERAE LRTSPQWLGPQIEVIRNATLSIEREINSVNDNPLIDVSRNKALHGRNFQGTPIGVSMDNT 316
SP|P45729|PAL3_PETCR LRTSPQWLGPQIEVIRSSTKMIEREINSVNDNPLIDVSRNKAIHGGNFQGSPIGVSMDNT 414
SP|P45726|PALY_CAMSI LRTSPQWLGPLIEVIRSSTKSIEREINSVNDNPLINVSRNKALHGGNFQGTPIGVSMDNT 410
SP|Q42609|PALY_BROFI LRTSPQWLGPQIEVIRAATKSIEREINSVNDNPLIDVSRNKALHGGNFQGTPIGVSMDNT 400
SP|O64963|PAL1_PRUAV LRTSPQWLGPQIEVIRYSTKSIEREIDSVNDNPLIDVSRNKALHGGNFQGTPIGVSMDNT 413
              ********** ***** .*  *****.********.******.** ****.*********

SP|P45724|PAL2_ARATH RLAIAAIGKLMFAQFSELVNDFYNNGLPSNLTASSNPSLDYGFKGAEIAMASYCSELQYL 473
SP|P45727|PALY_PERAE RLAIAAIGKLMFAQFSELVNDFYNNGLPSNLSGGRNPSLDYGFKGAEIAMAAYCSELQFL 376
SP|P45729|PAL3_PETCR RLAIAAIGKLMFAQFSELVNDFYNNGLPSNLSGGRNPSLDYGFKGAEIAMASYCSELQFL 474
SP|P45726|PALY_CAMSI RLAVASIGKLMFAQFSELVNDFYNNGLPSNLSGGRNPSLDYGFKGAEIAMAAYCSELQFL 470
SP|Q42609|PALY_BROFI RLAIAAIGKLMFAQFSELVNDFYNNGLPSNLSSGRNPSLDYGFKGAEIAMASYCSELQAL 460
SP|O64963|PAL1_PRUAV RLAIASIGKLMFAQFSELVNDFYNNGLPSNLSGGRNPSLDYGFKGAEIAMASYCSELQFL 473
              ***.*.************************...  ***************.****** *

SP|P45724|PAL2_ARATH ANPVTSHVQSAEQHNQDVNSLGLISSRKTSEAVDILKLMSTTFLVGICQAVDLRHLEENL 533
SP|P45727|PALY_PERAE ANPVTNHVQSAEQHNQDVNSLGLISSRKTAEAVEILKLMSSTFLVGLCQAIDLRHLEENL 436
```

```
SP|P45729|PAL3_PETCR ANPVTNHVQSAEQHNQDVNSLGLISSRKTSEAVEILKLMSTTFLVGLCQAIDLRHLEENL 534
SP|P45726|PALY_CAMSI ANPVTNHVQSAEQHNQDVNSLGLISSRKTAEAVDILKLMSSTYLVALCQAVDLRHFEENL 530
SP|Q42609|PALY_BROFI ANPVTNHVQSAEQHNQDVNSLGLISSRKTAEAVDILKLMSTTFLVGLCQAVDLRHLEENL 520
SP|O64963|PAL1_PRUAV ANPVTNHVQSAEQHNQDVNSLGLISSRKTAEAVDILKLMSSTFLVALCQAIDLRHLEENL 533
                     *****.**********************.***.******.*.** .***.****.****

SP|P45724|PAL2_ARATH RQTVKNTVSQVAKKVLTTGINGELHPSRFCEKDLLKVVDREQVFTYVDDPCSATYPLMQR 593
SP|P45727|PALY_PERAE KSTVKNTVSQVAKRVLTIGVNGELHPSRFCEKDLIKVVDGEHLFAYIDDPCSCTYPLMQK 496
SP|P45729|PAL3_PETCR KSTVKNTVSQVAKRVLTMGVNGELHPSRFCEKDLLRVVDREYIFAYIDDPCSATYPLMQK 594
SP|P45726|PALY_CAMSI RNTVKSTVSQVAKRVLTMGVNGELHPSRFCEKDLLRVVDREYIFAYIDDPCSATYPLMQK 590
SP|Q42609|PALY_BROFI KNAVKNTVSQVAKRVLTMGVNGELHPSRFCEKDLIKVIDREYVFAYADDPCSSTYPLMQK 580
SP|O64963|PAL1_PRUAV RNTVKNTVSQVAKRTLTTGVNGELHPSRFCEKDLLKVVDREYVFAYIDDPCSATYPLMQK 593
                     :.:**.*******.:** *.**************.:*.* * .*.* *****.******:

SP|P45724|PAL2_ARATH LRQVIVDHALSNGETEKNAVTSIFQKIGAFEEELKAVLPKEVEAARAAYGNGTAPIPNRI 653
SP|P45727|PALY_PERAE LRQVLVEHALINGEKEKDSSTSIFQKIGAFEEELKTHLPKEVESARIELERGNSAIPNRI 556
SP|P45729|PAL3_PETCR LRETLVEHALNNGDKERNLSTSIFQKIAAFEDELKALLPKEVETARAALESGNPAIPNRI 654
SP|P45726|PALY_CAMSI LRQVLVEHALKNGESEKNLSTSIFQKIRAFEEEIKTLLPKEVESTRAAIENGNSAIPNRI 650
SP|Q42609|PALY_BROFI LRAVIVEHALNNGVKEKDSNTSIFQKISSFENELKAALPKEVEAARAEFENGSPAIENRI 640
SP|O64963|PAL1_PRUAV LRQVLVEHALTNGENEKNASTSIFQKIVAFEEELKVLLPKEVDSARAALDSGSAGVPNRI 653
                     ** .:*.*** ** .*.::  ******* .**.*.*. *****.::*     *. : ***

SP|P45724|PAL2_ARATH KECRSYPLYRFVREELGTKLLTGEKVVSPGEEFDKVFTAMCEGKLIDPLMDCLKEWNGAP 713
SP|P45727|PALY_PERAE KECRSYPLYKFVREELKTSLLTGEKVRSPGEEFDKVFSAICQGKVIDPLLECLREWNGAP 616
SP|P45729|PAL3_PETCR KECRSYPLYKFVREELGTEYLTGEKVRSPGEEFEKVFTAMSKGEIIDPLLECLESWNGAP 714
SP|P45726|PALY_CAMSI KECRSYPLYKFVREELGTELLTGEKVRSPGEEFDKVFTALCKGEMIDPLMDCLKEWNGAP 710
SP|Q42609|PALY_BROFI KDCRSYPLYKFVK-EVGSGFLTGEKVVSPGEEFDKVFNAICEGKAIDPMLDCLKEWNGAP 699
SP|O64963|PAL1_PRUAV TECRSYPLYKFVREELGAEYLTGEKVRSPGEECDKVFTAICEGKIIDPILDCLEGWNGAP 713
                     .:*******.**.*. :  ****** ***** .*** *. .:*. ***.::**. *****

SP|P45724|PAL2_ARATH IPIC 717
SP|P45727|PALY_PERAE IPIC 620
SP|P45729|PAL3_PETCR LPIC 718
SP|P45726|PALY_CAMSI LPIC 714
SP|Q42609|PALY_BROFI LPIC 703
SP|O64963|PAL1_PRUAV LPIC 717
                     :***
```

6.10Command line prompts

Cutadapt and filtering and trimming of raw reads- (Section 3.3.1.5)

```
Cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -O 3
Andrewschoice_R1_001.fastq.gz -o Andrewschoice_R1_noadap_001.fastq.gz

Cutadapt -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -O 3
Andrewschoice_R2_001.fastq.gz -o Andrewschoice_R2_noadap_001.fastq.gz

Cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -O 3 Carlton_R1_001.fastq.gz -
o Carlton_R1_noadap_001.fastq.gz

Cutadapt -a AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-O 3 Carlton_R2_001.fastq.gz -o
Carlton_R2_noadap_001.fastq.gz


perl IlluQC.pl -pe Carlton_R1_noadap_001.fastq Calrton_R2_noadap_001.fastq N A -l 70 -s
20 -o Carlton_illuQC_out


perl IlluQC.pl -pe Andrewschoice_R1_noadap_001.fastq
Andrewschoice_R2_noadap_001.fastq N A -l 70 -s 20 -o Andrewschoice_illuQC_out

perl TrimmingReads.pl -i Carlton_R1_noadap_001.fastq_filtered -irev
Calrton_R2_noadap_001.fastq_filtered -q 30 -n 20

perl TrimmingReads.pl -i Andrewschoice_R1_noadap_001.fastq_filtered -irev
Andrewschoice_R2_noadap_001.fastq_filtered -q 30 -n 20
```

6.10.1 Trinity assembly and Soapdenovo-Trans assembly –(Section 3.3.1.6)

```
perl Trinity.pl --seqType fq --JM 54G --CPU 16 --left
Carlton_R1_noadap_001_Asssss.fastq_filtered_trimmed --right
Carlton_R2_noadap_001_Asssss.fastq_filtered_trimmed  --output
Carlton_trimmed_trinity_out.fasta
```

```
./SOAPdenovo-Trans all -s config_file -o soap_out 1> soap_stdout
```

configuration file:

```
[LIB]

max_rd_len=100

avg_ins=215

reverse_seq=1

asm_flags=3

rank=1

q1=JaneD_GCCAAT_L006_R1_noadap_001_Asssss.fastq_filtered_trimmed.fq

q2=JaneD_GCCAAT_L006_R2_noadap_001_Asssss.fastq_filtered_trimmed.fq
```

6.10.2 Mapping Analysis –(Section 3.3.1.7)

Build a reference:

```
bowtie2-build contig_sings_MID7_2_forblast 454_reference_

bowtie2-build  Trinity.fasta carlton_reference_

bowtie2-build Soap_denovo_trans.fasta soap_reference
```

map Illumina reads:

```
bowtie2  454_reference_ -1 Carlton_forward_trim_filt.fastq -2
Carlton_reverse_trin_filt.fastq Carlton_aligned_454_reads.sam

bowtie2  454_reference_ -1 Andrewschoice_forward_trim_filt.fastq -2
andrewschoice_reverse_trin_filt.fastq andrewschoice_aligned_454_reads.sam

bowtie2  carlton_reference_ -1 Carlton_forward_trim_filt.fastq -2
Carlton_reverse_trin_filt.fastq Carlton_aligned_trin_reads.sam

bowtie2  carlton_reference_ -1 Andrewschoice_forward_trim_filt.fastq -2
andrewschoice_reverse_trin_filt.fastq andrewschoice_aligned_trin_reads.sam

bowtie2  soap_reference_ -1 Carlton_forward_trim_filt.fastq -2
Carlton_reverse_trin_filt.fastq Carlton_aligned_soap_reads.sam

bowtie2  soap_reference_ -1 Andrewschoice_forward_trim_filt.fastq -2
andrewschoice_reverse_trin_filt.fastq andrewschoice_aligned_soap_reads.sam
```

(The following steps were carried out on all three assemblies, only the 454 assembly is shown for reference)

Index references:

```
samtools faidx contig_sings_MID_2_for blast
```

convert to bam file:

```
samtools import contig_sings_MID7_2_forblast.fai Carlton_aligned_to_454_reads.sam
Carlton_aligned_to_454_reads.bam
samtools import contig_sings_MID7_2_forblast.fai
Andrewschoice_aligned_to_454_reads.sam Andrewschoice_aligned_to_454_reads.bam
```

sort bam files:

```
samtools sort Carlton_aligned_to_454_reads.bam
Carlton_aligned_to_454_reads.sorted.bam
samtools sort Andrewschoice_aligned_to_454_reads.bam
Andrewschoice_aligned_to_454_reads.sorted.bam
```

index sorted bam files:

```
samtools index Carlton_aligned_to_454_reads.sorted.bam
samtools index Andrewschoice_aligned_to_454_reads.sorted.bam
```

(the following steps were carried out for both varieties with reads mapped to all three assemblies using the same options. For reference, only the Trinity reference steps are shown)

Run the perl script:

```
perl coverageStatsSplitByChr_v2.pl -i Carlton_aligned_reads_to_trinity.sorted.bam.bam
>Carlton_map_to_trinity_coverage.out
perl /pub16/jpulman/My_scripts/coverageStatsSplitByChr_v2.pl -i
Andrewschoice_aligned_reads_to_trinity.sorted.bam.bam
>Andrewschoice_map_to_trinity_coverage.out
```

pull out % mapped (testsplit.pl script set to pull out column 4):

```
perl  testsplit.pl Carlton_map_to_trinity_coverage.out >%mapped_Carlton_trinity
perl testsplit.pl Andrewschoice_map_to_trinity_coverage.out
>%mapped_Andrewschoice_trinity
```

calculate average:

```
awk 'BEGIN{s=0;}{s+=$1;}END{print s/NR;}' mapped_%_Carlton_trinity
awk 'BEGIN{s=0;}{s+=$1;}END{print s/NR;}' mapped_%_Andrewschoice_trinity
```

pull out coverage (testsplit.pl set to pull out column 5) :

```
perl testsplit.pl Carlton_map_to_trinity_coverage.out >Carlton_mean_depth_trinity
perl testsplit.pl Andrewschoice_map_to_trinity_coverage.out
>Andrewschoice_mean_depth_trinity
```

calculate average:

```
awk 'BEGIN{s=0;}{s+=$1;}END{print s/NR;}' mean_depth_Carlton_trinity


awk 'BEGIN{s=0;}{s+=$1;}END{print s/NR;}' mean_depth_Andrewschoice_trinity
```

6.10.3 VarScan and pileup_parser.pl –(Section 3.3.2)

```
java -jar VarScan.v2.3.4.jar pileup2snp Carlton_to_454_pileup --min-coverage 20 --min-
reads2 100 --min-avg-qual 20 --p-value 0.05 >varscanoptions_Carlton_out

java -jar VarScan.v2.3.4.jar pileup2snp Andrews_choice_to_454_pileup --min-coverage 20
--min-reads2 100 --min-avg-qual 20 --p-value 0.05 >varscanoptions_Andrewschoice_out

java -jar VarScan.v2.3.4.jar pileup2snp Carlton_to_trinity_pileup --min-coverage 20 --min-
reads2 100 --min-avg-qual 20 --p-value 0.05 >varscanoptions_Carlton_trinity_out

java -jar VarScan.v2.3.4.jar pileup2snp Andrews_choice_to_trinity_pileup --min-coverage
20 --min-reads2 100 --min-avg-qual 20 --p-value 0.05
>varscanoptions_Andrewschoice_trinity_out

Perl pileup_parser.pl Carlton_to454_pileup Calrton_454_pileup_parse_out

Perl pileup_parser.pl Andrewschoice__to454_pileup Andrews
choice_454_pileup_parse_out

Perl pileup_parser.pl Carlton_trinity_pileup Calrton_pileup_trinity_parse_out

Perl pileup_parser.pl Andrewschoice_trinity_pileup Andrews
choice_pileup__trinityparse_out
```

6.10.4 BitSeq Analysis – (section 3.3.2)

(*All the steps shown in this section were carried out on both assemblies, for reference the steps are shown on a generic set of files)*

```
bowtie-build -f -o 2 -t 12 --ntoa reference.fasta reference_index
```

```
bowtie -q -v 3 -3 0 -p 4 -a -m 100 --sam reference_index -1 condition1_forward.fastq -2
condition1_reverse.fastq condition1_trin_bitseq2.sam
```

```
bowtie -q -v 3 -3 0 -p 4 -a -m 100 --sam reference_index -1 condition2_forward.fastq -2
condition2_reverse.fastq condition2_trin_bitseq.sam
```

Step1: Pre-processing

```
./parseAlignment condition1_trin_bitseq.sam -o condition1.prob --trSeqFile
reference.fasta --trInfoFile reference.fasta.tr --uniform --verbose
```

```
./parseAlignment condition_2_trin_bitseq.sam -o condition2.prob --trSeqFile
reference.fasta --trInfoFile reference.fasta.tr --uniform –verbose
```

Step2: Sampling

```
./estimateExpression condition1.prob -o condition1 --outType RPKM -p parameters1.txt -
t reference.fasta.tr -P 4
./estimateExpression conditon2.prob -o condition2 --outType RPKM -p parameters1.txt -t
reference.fasta.tr -P 4
./getVariance -o conditon1.mean condition1.rpkm
./getVariance -o condition2.mean condition2.rpkm
```

Step 3:

```
./getVariance --log –o both_to_Trinity_ref.Lmean condition1.rpkm  condition2.rpkm
```

Hyper parameters:

```
./estimateHyperPar --meanFile both_to_Trinity_ref.Lmean -o both_to_Trinity_ref.param
condition1.rpkm C condition2.rpkm
```

Step4: Condition specific expression

```
./estimateDE -o both_to_Trinity_ref -p both_to_Trinity_ref.param condition1.rpkm C
condition2.rpkm
```

```
awk 'NR>8' both_to_Trinity.pplr > both_to_Trinity _noheader.pplr
```

```
paste trin_contigs both_to_Trinity _noheader.pplr > both_to_Trinity _final.pplr
```

```
perl tabtospace.pl both_to_Trinity _final.pplr both_to_Trinity _notab.pplr
```

```
perl testsplit_bitseq.pl both_to_Trinity_notab.pplr trinity_up trinity_down
```

6.10.5 BMtagger –(Section 3.3.2)

```
./bmtool -d TREP.complete.fas -o TREP.bitmask -A 0 -w 18
./srprism mkindex -i TREP.complete.fas -o TREP.srprism -M 7168
makeblastdb -in TREP.complete.fas -dbtype nucl
./bmtagger.sh -b TREP.bitmask -x TREP.srprism -T tmpjane -q1 -1
Carlton_forward.fastq -2 Carlton_reverse.fastq -o Carlton_bmtagger_output
```

```
perl extract_nonhuman_reads_fastq.pl Carlton_bmtagger_output
Carlton_trimmed_filtered_forward.fastq Carlton_trimmed_filtered_reverse.fastq
```

```
perl extract_nonhuman_reads_fastq.pl Andrews_choice_bmtagger_output
Andrews_choice_trimmed_filtered_forward.fastq
Andrews_choice_trimmed_filtered_reverse.fastq
```

6.10.6 TransposonPSI.pl –(Section 3.3.5.2)

```
perl transposonPSI.pl Carlton_R1_filt_trim.fasta nuc

perl transponsonPSI.pl Carlton_R2_filt_trim.fasta nuc

perl transposonPSI.pl Andrews_choice_R1_filt_trim.fasta nuc

perl transposonPSI.pl Andrews_choice_R2_filt_trim.fasta nuc
```

6.10.7 BLAST against TREP- (Section 3.3.5.3)

```
time pblastall Andrews_choice_forward.fasta -p blastx id TREP.complete.fas -F T -m 8 -b 1
-v 1 >JaneA_R1_blast_out.gz
time pblastall Andrews_choice_reverse.fasta -p blastx id TREP.complete.fas -F T -m 8 -b 1
-v 1 >JaneA_R2_blast_out.gz

time pblastall Carlton_forward.fasta -p blastx id TREP.complete.fas -F T -m 8 -b 1 -v 1
>JaneD_R1_blast_out.gz
time pblastall Carlton_reverse.fasta -p blastx id TREP.complete.fas -F T -m 8 -b 1 -v 1
>JaneD_R2_blast_out.gz

perl extract_nonhuman_reads_fastq.pl  blast_against_TREP_e5_ids Carlton_forward.fastq
Carlton_reverse.fastq
```

```
perl extract_nonhuman_reads_fastq.pl  blast_against_TREP_e5_ids
Andrews_choice_forward.fastq Andrews_choice_reverse.fastq
```

6.10.8 Trinity re-assembly- (Section 3.3.6)

```
perl Trinity.pl --seqType fq --JM 54G --left Carlton_forward_BM_trimmed_filtered.fastq --
right Calrton_reverse_BM_trimmed_filtered.fastq --CPU 6
```

6.10.9 BLAST against genes of interest database- (Section 4.2.3)

```
time pblastall 454_reference.fasta -p blastx -d predicted_genes.fasta -F T -m 8 -b 1 -v 1
>Predicted_gene_454_blast_out.gz

gunzip Predicted_gene_454_blast_out.gz

perl remove_low_scoring_blast.pl Predicted_gene_454_blast_out_rlsb

cat Predicted_gene_454_blast_out_rlsb  | firstm8hsp.sh >
Predicted_gene_454_blast_out_rlsb_nhsp
```

6.11 References

(1) Heinrich, M., and Teoh, H. L. (2004) Galanthamine from snowdrop - the development of a modern drug against Alzheimer's disease from local Caucasian knowledge. *Journal of Ethnopharmacology 92*, 147–162.

(2) Xiao, M., Zhang, Y., Chen, X., Lee, E.-J., Barber, C. J. S., Chakrabarty, R., Desgagne-Penix, I., Haslam, T. M., Kim, Y.-B., Liu, E., MacNevin, G., Masada-Atsumi, S., Reed, D. W., Stout, J. M., Zerbe, P., Zhang, Y., Bohlmann, J., Covello, P. S., De Luca, V., Page, J. E., Ro, D.-K., Martin, V. J. J., Facchini, P. J., and Sensen, C. W. (2013) Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *Journal of Biotechnology 166*, 122–134.

(3) Ziegler, J., Facchini, P. J., Geissler, R., Schmidt, J., Ammer, C., Kramell, R., Voigtlaender, S., Gesell, A., Pienkny, S., and Brandt, W. (2009) Evolution of morphine biosynthesis in opium poppy. *Phytochemistry 70*, 1696–1707.

(4) Eichhorn, J., Takada, T., Kita, Y., and Zenk, M. H. (1998) Biosynthesis of the Amaryllidaceae alkaloid galanthamine. *Phytochemistry*.

(5) Guo, X., Li, Y., Li, C., Luo, H., Wang, L., Qian, J., Luo, X., Xiang, L., Song, J., Sun, C., Xu, H., Yao, H., and Chen, S. (2013) Analysis of theDendrobium officinaletranscriptome reveals putative alkaloid biosynthetic genes and genetic markers. *Gene 527*, 131–138.

(6) Facchini, P. J., Bohlmann, J., Covello, P. S., De Luca, V., Mahadevan, R., Page, J. E., Ro, D.-K., Sensen, C. W., Storms, R., and Martin, V. J. J. (2012) Synthetic biosystems for the production of high-value plant metabolites. *Trends in Biotechnology 30*, 127–131.

(7) Facchini, P. J., Huber-Allanach, K. L., and Tari, L. W. (2000) Plant aromatic L-amino acid decarboxylases: evolution, biochemistry, regulation, and metabolic engineering applications. *Phytochemistry 54*, 121–138.

(8) Beaudoin, G. A. W., and Facchini, P. J. (2014) Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta 240*, 19–32.

(9) McGinn, S., and Gut, I. G. (2013) DNA sequencing – spanning the generations. *New Biotechnology 30*, 366–372.

(10) Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012) Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology 2012*, 1–11.

(11) Bräutigam, A., and Gowik, U. (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology 12*, 831–841.

(12) Hornett, E. A., and Wheat, C. W. (2012) Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics 13*, 361.

(13) Martin, L. B. B., Fei, Z., Giovannoni, J. J., and Rose, J. K. C. (2013) Catalyzing plant science research with RNA-seq. *Front Plant Sci 4*, 66.

(14) Liu, S., Li, W., Wu, Y., Chen, C., and Lei, J. (2013) De novo transcriptome assembly in chili pepper (Capsicum frutescens) to identify genes involved in the biosynthesis of capsaicinoids. *PLoS ONE 8*, e48156–e48156.

(15) Desgagne-Penix, I., Farrow, S. C., Cram, D., Nowak, J., and Facchini, P. J. (2012) Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Mol Biol 79*, 295–313.

(16) Liu, S., Chen, C., Chen, G., Cao, B., Chen, Q., and Lei, J. (2012) RNA-sequencing tag profiling of the placenta and pericarp of pungent pepper provides robust candidates contributing to capsaicinoid biosynthesis. *Plant Cell Tiss Organ Cult 110*, 111–121.

(17) Ikezawa, N., Iwasa, K., and Sato, F. (2008) CYP719A subfamily of cytochrome P450 oxygenases and isoquinoline alkaloid biosynthesis in Eschscholzia californica. *Plant Cell Rep 28*, 123–133.

(18) Lee, E. J., and Facchini, P. (2010) Norcoclaurine Synthase Is a Member of the Pathogenesis-Related 10/Bet v1 Protein Family. *The Plant Cell 22*, 3489–3503.

(19) Park, N. I., Choi, I. Y., Choi, B.-S., Kim, Y. S., Lee, M. Y., and Park, S.-U. (2014) EST sequencing and gene expression profiling in Scutellaria baicalensis.

(20) Xiao, Y., Zhang, L., Gao, S., Saechao, S., Di, P., Chen, J., and Chen, W. (2011) The c4h, tat, hppr and hppd Genes Prompted Engineering of Rosmarinic Acid Biosynthetic Pathway in Salvia miltiorrhiza Hairy Root Cultures. *PLoS ONE* (Uversky, V. N., Ed.) *6*, e29713.

(21) Curry, J., Aluru, M., Mendoza, M., Nevarez, J., and Melendrez, M. (1999) Transcripts for

possible capsaicinoid biosynthetic genes are differentially accumulated in pungent and non-pungent< i> Capsicum</i>< i> spp</i>. *Plant Science*.

(22) Gupta, P. K., Roy, J. K., and Prasad, M. (2001) Single nucleotide polymorphisms (SNPs): a new paradigm in molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Current Science*.

(23) Andersen, J. R., and Lübberstedt, T. (2003) Functional markers in plants. *Trends in Plant Science 8*, 554–560.

(24) Van, K., Kim, D. H., Shin, J. H., and Lee, S.-H. (2011) Genomics of plant genetic resources: past, present and future. *Plant Genet. Resour. 9*, 155–158.

(25) Deschamps, S., and Campbell, M. A. (2009) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol Breeding 25*, 553–570.

(26) Hanks, G. R. (2002) Chapter 1, 4, in *DNA sequencing – spanning the generations* (Hanks, G. R., Ed.). Taylor and Francis, London.

(27) Ronsted, N., Savolainen, V., Molgaard, P., and Jager, A. K. (2008) Phylogenetic selection of Narcissus species for drug discovery. *Biochemical Systematics and Ecology 36*.

(28) Graham, S. W., and Barrett, S. C. H. (2004) Phylogenetic reconstruction of the evolution of stylar polymorphisms in Narcissus (Amaryllidaceae). *Am. J. Bot. 91*, 1007–1021.

(29) Medrano, M., López-Perea, E., and Herrera, C. M. (2014) Population Genetics Methods Applied to a Species Delimitation Problem: Endemic Trumpet Daffodils (NarcissusSection Pseudonarcissi) from the Southern Iberian Peninsula. *International Journal of Plant Sciences 175*, 501–517.

(30) Zonneveld, B. J. M. (2008) The systematic value of nuclear DNA content for all species of Narcissus L. (Amaryllidaceae). *Plant Syst Evol 275*, 109–132.

(31) Brandham, P. E., and West, J. P. (1993) Correlation between nuclear DNA values and differing optimal ploidy levels inNarcissus, Hyacinthus andTulipa cultivars. *Genetica 90*, 1–8.

(32) Takos, A., and Rook, F. (2013) Towards a Molecular Understanding of the Biosynthesis of Amaryllidaceae Alkaloids in Support of Their Expanding Medical Use. *IJMS 14*, 11713–11741.

(33) Jin, Z. (2009) Amaryllidaceae and Sceletium alkaloids. *Nat Prod Rep 26*, 363–381.

(34) Jin, Z. (2005) Amaryllidaceae and Sceletium alkaloids. *Nat Prod Rep 22*, 111–126.

(35) Berkov, S., Georgieva, L., Kondakova, V., Viladomat, F., Bastida, J., Atanassov, A., and Codina, C. (2013) The geographic isolation of Leucojum aestivum populations leads to divergation of alkaloid biosynthesis. *Biochemical Systematics and Ecology 46*, 152–161.

(36) Plaitakis, A., and Duvoisin, R. C. (1983) HOMERS MOLY IDENTIFIED AS GALANTHUS-NIVALIS L - PHYSIOLOGIC ANTIDOTE TO STRAMONIUM POISONING. *Clinical Neuropharmacology 6*, 1–5.

(37) Bastida, J., Lavilla, R., and Viladomat, F. (2006) CHEMICAL AND BIOLOGICAL ASPECTS OF NARCISSUS ALKALOIDS. *Alkaloids: Chemistry and Biology, Vol 63 63*, 87–179.

(38) Ogita, S. U. H. Y. Y. K. N. S. H. (2009) NICE implementation uptake report: Donepezil, galathamine, rivastigmine and memantine for the treatment of Alzheimer's disease. *Nature 423*, 823.

(39) Proskurina, N. F., and Yakovleva, A. P. (1952) Alkaloids of Galanthus woronowi. II Isolation of a new alkaloid. *Zurnal Obshchei Khimii* 1899–1902.

(40) Mashkovskii, M. D. (1955) Effects of galanthamine on the acetylcholine sensitivity of skeletal musculature. Farmakol Toksikol.

(41) McNulty, J., Nair, J. J., Singh, M., Crankshaw, D. J., and Holloway, A. C. (2010) Potent and selective inhibition of human cytochrome P450 3A4 by seco-pancratistatin structural analogs. *Bioorganic & Medicinal Chemistry Letters 20*, 2335–2339.

(42) McNulty, J., Nair, J. J., Little, J. R. L., Brennan, J. D., and Bastida, J. (2010) Structure-activity studies on acetylcholinesterase inhibition in the lycorine series of Amaryllidaceae alkaloids. *Bioorganic & Medicinal Chemistry Letters 20*, 5290–5294.

(43) Lamoral-Theys, D., Decaestecker, C., Mathieu, V., Dubois, J., Kornienko, A., Kiss, R., Evidente, A., and Pottier, L. (2010) Lycorine and its Derivatives for Anticancer Drug Design. *Mini-Reviews in Medicinal Chemistry 10*, 41–50.

(44) Cedron, J. C., Gutierrez, D., Flores, N., Ravelo, A. G., and Estevez-Braun, A. (2010) Synthesis and antiplasmodial activity of lycorine derivatives. *Bioorganic & Medicinal Chemistry 18*, 4694–4701.

(45) Hwang, Y.-C., Chu, J. J.-H., Yang, P. L., Chen, W., and Yates, M. V. (2008) Rapid identification of inhibitors that interfere with poliovirus replication using a cell-based assay. *Antiviral*

Research 77, 232–236.
(46) Li, S. Y., Chen, C., Zhang, H. Q., Guo, H. Y., Wang, H., Wang, L., Zhang, X., Hua, S. N., Yu, J., Xiao, P. G., Li, R. S., and Tan, X. H. (2005) Identification of natural compounds with antiviral activities against SARS-associated coronavirus. *Antiviral Research 67*, 18–23.
(47) Wang, P., Li, L. F., Wang, Q. Y., Shang, L. Q., Shi, P. Y., and Yin, Z. (2014) Anti-Dengue-Virus Activity and Structure–Activity Relationship Studies of Lycorine Derivatives. *ChemMedChem 9*, 1522–1533.
(48) Shen, J. W., Ren, W., and Wang, X. L. (2014) Lycorine: A Potential Broad-Spectrum Agent Against Crop Pathogenic Fungi. *Journal of microbiology and …*.
(49) Van Goietsenoven, G., Andolfi, A., Lallemand, B., Cimmino, A., Lamoral-Theys, D., Gras, T., Abou-Donia, A., Dubois, J., Lefranc, F., Mathieu, V., Kornienko, A., Kiss, R., and Evidente, A. (2010) Amaryllidaceae Alkaloids Belonging to Different Structural Subgroups Display Activity against Apoptosis-Resistant Cancer Cells. *Journal of Natural Products 73*, 1223–1227.
(50) Van Goietsenoven, G., Mathieu, V., Lefranc, F., Kornienko, A., Evidente, A., and Kiss, R. (2013) Narciclasine as well as other Amaryllidaceae Isocarbostyrils are Promising GTP-ase Targeting Agents against Brain Cancers. *Medicinal Research Reviews 33*, 439–455.
(51) Loizzo, M. R., Tundis, R., Menichini, F., and Menichini, F. (2008) Natural products and their derivatives as cholinesterase inhibitors in the treatment of neurodegenerative disorders: An update. *Curr. Med. Chem. 15*, 1209–1228.
(52) Howes, M. J. R., Perry, N. S. L., and Houghton, P. J. (2003) Plants with traditional uses and activities, relevant to the management of Alzheimer's disease and other cognitive disorders. *Phytotherapy Research 17*.
(53) Kane, M., and Cook, L. (2013) Dementia 2013: The hidden voice of loneliness. London: Alzheimers Society.
(54) Suh, G.-H., Wimo, A., Gauthier, S., O'Connor, D., Ikeda, M., Homma, A., Dominguez, J., Yang, B.-M., and Int Psychogeriatric, A. (2009) International price comparisons of Alzheimer's drugs: a way to close the affordability gap. *International Psychogeriatrics 21*, 1116–1126.
(55) Selkoe, D. J. (2001) Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews 81*, 741–766.
(56) Schrattenholz, A., Pereira, E. F. R., Roth, U., Weber, K. H., Albuquerque, E. X., and Maelicke, A. (1996) Agonist responses of neuronal nicotinic acetylcholine receptors are potentiated by a novel class of allosterically acting ligands. *Molecular Pharmacology 49*, 1–6.
(57) Storch, A., Schrattenholz, A., Cooper, J. C., Ghani, E. M. A., Gutbrod, O., Weber, K. H., Reinhardt, S., Lobron, C., Hermsen, B., Soskic, V., Pereira, E. F. R., Albuquerque, E. X., Methfessel, C., and Maelicke, A. (1995) PHYSOSTIGMINE, GALANTHAMINE AND CODEINE ACT AS NONCOMPETITIVE NICOTINIC RECEPTOR AGONISTS ON CLONAL RAT PHEOCHROMOCYTOMA CELLS. *European Journal of Pharmacology-Molecular Pharmacology Section 290*, 207–219.
(58) Guillou, C., Beunard, J. L., Gras, E., and Thal, C. (2001) An Efficient Total Synthesis of (±)-Galanthamine. *Angewandte Chemie*.
(59) Liscombe, D. K., and Facchini, P. J. (2008) Evolutionary and cellular webs in benzylisoquinoline alkaloid biosynthesis. *Current Opinion in Biotechnology 19*, 173–180.
(60) Barton, D. H. R., and Kirby, G. W. (1962) PHENOL OXIDATION AND BIOSYNTHESIS .5. SYNTHESIS OF GALANTHAMINE. *Journal of the Chemical Society* 806–&.
(61) Tanimoto, H., Kato, T., and Chida, N. (2007) Total synthesis of (+)-galanthamine starting from D-glucose. *Tetrahedron Letters 48*.
(62) Frick, S., Kramell, R., and Kutchan, T. M. (2007) Metabolic engineering with a morphine biosynthetic P450 in opium poppy surpasses breeding. *Metab Eng 9*, 169–176.
(63) Barton, D. H. R. A. C. T. (1957) Festschrift arthur stoll. Birkhauser, Basel.
(64) Fuganti, C. (1969) BIOSYNTHESIS OF HAEMANTHAMINE-STEREOCHEMISTRY OF PROTONATION AT C-4. CHIMICA & L INDUSTRIA.
(65) Bhandark Jg, and Kirby, G. W. (1970) STRUCTURE AND BIOSYNTHESIS OF CHLIDANTHINE. *J. Chem. Soc., C* 1224–&.
(66) Zenk, M. H., Gerardy, R., and Stadler, R. (1989) Phenol oxidative coupling of benzylisoquinoline alkaloids is catalysed by regio- and stereo-selective cytochrome P-450 linked plant enzymes: salutaridine and berbamunine. *J. Chem. Soc., Chem. Commun.* 1725.
(67) Kraus, P. F. X., and Kutchan, T. M. (1995) MOLECULAR-CLONING AND HETEROLOGOUS EXPRESSION OF A CDNA-ENCODING BERBAMUNINE SYNTHASE, A C-O PHENOL-COUPLING CYTOCHROME-P450 FROM THE HIGHER-PLANT BERBERIS-STOLONIFERA. *PNAS 92*, 2071–

2075.
(68) Nasreen, A., Rueffer, M., and Zenk, M. H. (1996) Cytochrome P-450-dependent formation of isoandrocymbine from autumnaline in colchicine biosynthesis. *Tetrahedron Letters 37*, 8161–8164.
(69) Schuler, M. A., and Werck-Reichhart, D. (2003) Functional genomics of P450s. *Annu. Rev. Plant Biol. 54*, 629–667.
(70) Gerardy, R., and Zenk, M. H. (1993) Purification and characterization of salutaridine: NADPH 7-oxidoreductase from Papaver somniferum. *Phytochemistry*.
(71) Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature 408*, 796–815.
(72) Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchinson, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J. P., Miguel, T., Paszkowski, U., Zhang, S. P., Colbert, M., Sun, W. L., Chen, L. L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y. S., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). *Science 296*, 92–100.
(73) Yu, J., Hu, S. N., Wang, J., Wong, G., Li, S. G., Liu, B., Deng, Y. J., Dai, L., Zhou, Y., Zhang, X. Q., Cao, M. L., Liu, J., Sun, J. D., Tang, J. B., Chen, Y. J., Huang, X. B., Lin, W., Ye, C., Tong, W., Cong, L. J., Geng, J. N., Han, Y. J., Li, L., Li, W., Hu, G. Q., Huang, X. G., Li, W. J., Li, J., Liu, Z. W., Liu, J. P., Qi, Q. H., Liu, J. S., Li, T., Wang, X. G., Lu, H., Wu, T. T., Zhu, M., Ni, P. X., Han, H., Dong, W., Ren, X. Y., Feng, X. L., Cui, P., Li, X. R., Wang, H., Xu, X., Zhai, W. X., Xu, Z., Zhang, J. S., He, S. J., Zhang, J. G., Xu, J. C., Zhang, K. L., Zheng, X. W., Dong, J. H., Zeng, W. Y., Tao, L., Ye, J., Tan, J., Ren, X. D., Chen, X. W., He, J., Liu, D. F., Tian, W., Tian, C. G., Xia, H. G., Bao, Q. Y., Li, G., Gao, H., Cao, T., Zhao, W. M., Li, P., Chen, W., Wang, X. D., Zhang, Y., Hu, J. F., Liu, S., Yang, J., Zhang, G. Y., Xiong, Y. Q., Li, Z. J., Mao, L., Zhou, C. S., Zhu, Z., Chen, R. S., Hao, B. L., Zheng, W. M., Chen, S. Y., Guo, W., Li, G. J., Liu, S. Q., Tao, M., Zhu, L. H., Yuan, L. P., and Yang, H. M. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp indica). *Science 296*, 79–92.
(74) Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., Grigoriev, I. V., Lyons, E., Maher, C. A., Martis, M., Narechania, A., Otillar, R. P., Penning, B. W., Salamov, A. A., Wang, Y., Zhang, L., Carpita, N. C., Freeling, M., Gingle, A. R., Hash, C. T., Keller, B., Klein, P., Kresovich, S., McCann, M. C., Ming, R., Peterson, D. G., Mehboob-ur-Rahman, Ware, D., Westhoff, P., Mayer, K. F. X., Messing, J., and Rokhsar, D. S. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature 457*, 551–556.
(75) Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature 463*, 178–183.
(76) McGinn, S., and Gut, I. G. (2013) DNA sequencing – spanning the generations. *New Biotechnology 30*, 366–372.
(77) Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010) The next-generation sequencing technology and application. *Protein Cell 1*, 520–536.
(78) Rothberg, J. M., and Leamon, J. H. (2008) The development and impact of 454 sequencing. *Nat Biotechnol 26*.
(79) Roche. (2014) 454 sequencing. *454.com*.
(80) Illumina. (2010) Illumina Sequencing Technology.
(81) Mardis, E. R. (2013) Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry 6*, 287–303.
(82) Thudi, M. M., Li, Y. Y., Jackson, S. A. S., May, G. D. G., and Varshney, R. K. R. (2012) Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomic Proteomic 11*, 3–11.
(83) Schliesky, S., Gowik, U., Weber, A. P. M., and Bräutigam, A. (2012) RNA-Seq Assembly - Are

We There Yet? *Front Plant Sci 3*, 220.
(84) Strickler, S. R., Bombarely, A., and Mueller, L. A. (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *Am. J. Bot. 99*, 257–266.
(85) Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*.
(86) Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics 5*, 433–438.
(87) Bennett, S. T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics 6*, 373–382.
(88) Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X. X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science 309*, 1728–1732.
(89) Morozova, O., and Marra, M. A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics 92*, 255–264.
(90) Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G. K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010) The sequence and de novo assembly of the giant panda genome (vol 463, pg 311, 2010). *Nature 463*, 1106–1106.
(91) Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome biology 12*, R112–15.
(92) Weber, A. P. M., Weber, K. L., Carr, K., Wilkerson, C., and Ohlrogge, J. B. (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *PLANT PHYSIOLOGY 144*, 32–42.
(93) Der, J. P., Barker, M. S., Wickett, N. J., dePamphilis, C. W., and Wolf, P. G. (2011) De novo characterization of the gametophyte transcriptome in bracken fern, Pteridium aquilinum. *BMC Genomics 12*, 99–99.
(94) Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F., and Myburg, A. A. (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. *BMC Genomics 11*, 681.
(95) Sun, X., Zhou, S., Meng, F., and Liu, S. (2012) De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing. *Plant Cell Rep 31*, 1823–1828.
(96) Franssen, S. U., Shrestha, R. P., Brutigam, A., Bornberg-Bauer, E., and Weber, A. P. (2011) Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing. *BMC Genomics*
12
, 227–227.
(97) Barakat, A., DiLoreto, D. S., Zhang, Y., Smith, C., Baier, K., Powell, W. A., Wheeler, N., Sederoff, R., and Carlson, J. E. (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. *BMC Plant Biol 9*, 51.
(98) Garg, R., Patel, R. K., Tyagi, A. K., and Jain, M. (2011) De novo assembly of chickpea

transcriptome using short reads for gene discovery and marker identification. *DNA research*.

(99) Wong, G. K.-S., Deyholos, M., and Zhang, Y. 1000 Plants. *httpssites.google.comaualberta.caonekphome*.

(100) Dai, N., Cohen, S., Portnoy, V., Tzuri, G., and Harel-Beja, R. (2011) Metabolism of soluble sugars in developing melon fruit: a global transcriptional view of the metabolic transition to sucrose accumulation. *Plant molecular …*.

(101) Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., Giuliano, G., Chiusano, M. L., Baldoni, L., and Perrotta, G. (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics 10*, 399.

(102) Bräutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K. L., Carr, K. M., Gowik, U., Maß, J., Lercher, M. J., Westhoff, P., Hibberd, J. M., and Weber, A. P. M. (2011) An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *PLANT PHYSIOLOGY 155*, 142–156.

(103) Sun, J. Q., Jiang, H. L., and Li, C. Y. (2011) Systemin/jasmonate-mediated systemic defense signaling in tomato. *Molecular plant*.

(104) Staff, B.-I. W. (2013) Six Years After Acquisition, Roche Quietly Shutters 454. *http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shutters-454.html*.

(105) Biosciences, P. (2013) SMRT Technology.

(106) (null), O. N. T. (2014) DNA: An introduction to nanopore sequencing. *https://www.nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/dna-an-introduction-to-nanopore-sequencing*.

(107) Timp, W., Mirsaidov, U. M., Deqiang Wang, Comer, J., Aksimentiev, A., and Timp, G. (2010) Nanopore Sequencing: Electrical Measurements of the Code of Life. *IEEE Trans. Nanotechnology 9*, 281–294.

(108) Larsson, C., Grundberg, I., Söderberg, O., and Nilsson, M. (2010) In situ detection and genotyping of individual mRNA molecules. *Nat Methods 7*, 395–397.

(109) Hirsch, C. N., and Robin Buell, C. (2013) Tapping the Promise of Genomics in Species with Complex, Nonmodel Genomes. *Annu. Rev. Plant Biol. 64*, 89–110.

(110) Góngora-Castillo, E., Childs, K. L., Fedewa, G., Hamilton, J. P., Liscombe, D. K., Magallanes-Lundback, M., Mandadi, K. K., Nims, E., Runguphan, W., Vaillancourt, B., Varbanova-Herde, M., DellaPenna, D., McKnight, T. D., O'Connor, S., and Buell, C. R. (2012) Development of Transcriptomic Resources for Interrogating the Biosynthesis of Monoterpene Indole Alkaloids in Medicinal Plant Species. *PLoS ONE* (Mariño-Ramírez, L., Ed.) *7*, e52506.

(111) Sangwan, R. S., Tripathi, S., Singh, J., Narnoliya, L. K., and Sangwan, N. S. (2013) De novo sequencing and assembly of Centella asiatica leaf transcriptome for mapping of structural, functional and regulatory genes with special reference to secondary metabolism. *Gene 525*, 58–76.

(112) Desgagne-Penix, I., Khan, M. F., Schriemer, D. C., Cram, D., Nowak, J., and Facchini, P. J. (2010) Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. *BMC Plant Biol 10*, 252.

(113) Hagel, J. M., and Facchini, P. J. (2010) Dioxygenases catalyze the O-demethylation steps of morphine biosynthesis in opium poppy. *Nat Chem Biol 6*, 273–275.

(114) Lee, E.-J., and Facchini, P. J. (2011) Tyrosine aminotransferase contributes to benzylisoquinoline alkaloid biosynthesis in opium poppy. *PLANT PHYSIOLOGY 157*, 1067–1078.

(115) Zerbe, P., Chiang, A., Yuen, M., Hamberger, B., Hamberger, B., Draper, J. A., Britton, R., and Bohlmann, J. (2012) Bifunctional cis-abienol synthase from Abies balsamea discovered by transcriptome sequencing and its implications for diterpenoid fragrance production. *J. Biol. Chem. 287*, 12121–12131.

(116) Nguyen, T. D., MacNevin, G., and Ro, D. K. (2012) De novo synthesis of high-value plant sesquiterpenoids in yeast. *Methods Enzymol*.

(117) B, P., DP, D., T, M., C, W., HT, S., and D-K, R. (2012) Identification and characterization of a kunzeaol synthase from Thapsia garganica: implications for the biosynthesis of the pharmaceutical thapsigargin. *Biochemical Journal 448*, 261–271.

(118) Woldemariam, M. G., Baldwin, I. T., and Galis, I. (2011) Transcriptional regulation of plant inducible defenses against herbivores: a mini-review. *Journal of Plant Interactions 6*, 113–119.

(119) Zhao, N., Wang, G., Norris, A., Chen, X., and Chen, F. (2013) Studying Plant Secondary Metabolism in the Age of Genomics. *Critical Reviews in Plant Sciences 32*, 369–382.

(120) Fridman, E., Wang, J., Iijima, Y., Froehlich, J. E., Gang, D. R., Ohlrogge, J., and Pichersky, E. (2005) Metabolic, genomic, and biochemical analyses of glandular trichomes from the wild tomato species Lycopersicon hirsutum identify a key enzyme in the biosynthesis of methylketones. *The Plant Cell 17*, 1252–1267.
(121) Liscombe, D. K., Ziegler, J. R., Schmidt, J. R., Ammer, C., and Facchini, P. J. (2009) Targeted metabolite and transcript profiling for elucidating enzyme function: isolation of novel N-methyltransferases from three benzylisoquinoline alkaloid-producing species. *The Plant Journal 60*, 729–743.
(122) Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet 43*, 956–963.
(123) Shu, Q. Y., Wu, D. X., Xia, Y. W., Gao, M. W., and Ayres, N. M. (1999) Microsatellites polymorphism on the waxy gene locus and their relationship to amylose content in indica and japonica rice, Oryza sativa L. Acta Genet Sin.
(124) Garland, S. H., Lewin, L. G., Blakeney, A. B., and Henry, R. J. (2000) Microsatellite and SNP markers for sd-1 in rice.
(125) Zhu, Y. L., Song, Q. J., Hyten, D. L., Van Tassell, C. P., Matukumalli, L. K., Grimm, D. R., Hyatt, S. M., Fickus, E. W., Young, N. D., and Cregan, P. B. (2003) Single-nucleotide polymorphisms in soybean. *Genetics 163*, 1123–1134.
(126) Alcala, J., Giovannoni, J. J., Pike, L. M., and Reddy, A. S. (1997) Application of Genetic Bit Analysis (GBATM) for allelic selection in plant breeding. *Mol Breeding 3*, 495–502.
(127) zhang, J., Ruhlman, T. A., Mower, J. P., and Jansen, R. K. (2013) Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. *BMC Plant Biol 13*, 228–228.
(128) Clarke, K., Yang, Y., Marsh, R., Xie, L., and Zhang, K. K. (2013) Comparative analysis of de novo transcriptome assembly. *Sci. China Life Sci. 56*, 156–162.
(129) Chevreux, B., Wetter, T., and Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. *German Conference on Bioinformatics*.
(130) Mullikin, J. C. (2002) The Phusion Assembler. *Genome Research 13*, 81–90.
(131) Huang, X. (1999) CAP3: A DNA Sequence Assembly Program. *Genome Research 9*, 868–877.
(132) Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol 29*, 644–652.
(133) Annadurai, R. S., Neethiraj, R., Jayakumar, V., Damodaran, A. C., Rao, S. N., Katta, M. A. V. S. K., Gopinathan, S., Sarma, S. P., Senthilkumar, V., Niranjan, V., Gopinath, A., and Mugasimangalam, R. C. (2013) De Novo Transcriptome Assembly (NGS) of Curcuma longa L. Rhizome Reveals Novel Transcripts Related to Anticancer and Antimalarial Terpenoids. *PLoS ONE* (Aggarwal, B. B., Ed.) *8*, e56217.
(134) Jaakola, L., Pirttila, A. M., Halonen, M., and Hohtola, A. (2001) Isolation of high quality RNA from bilberry (Vaccinium myrtillus L.) fruit. *Molecular Biotechnology 19*, 201–203.
(135) Dang, P. M., and Chen, C. Y. (2013) Modified method for combined DNA and RNA isolation from peanut and other oil seeds. *Mol Biol Rep 40*, 1563–1568.
(136) Chang, S., Puryear, J., and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter 11*, 113–116.
(137) life-technologies. (2008) Dynabeads mRNA purification kit for mRNA purification from total RNA preps. *lifetechnologies.com*.
(138) Invitrogen, life-technologies. RiboMinus Plant Kit for RNA-Seq. *httptools.lifetechnologies.comcontentsfsmanualsribominusplantman.pdf*.
(139) Ferland, D. L. H. (2011) cDNA Rapid Library Preparation Method Manual 1–12.
(140) Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Da Luo, Li, X., and Hao, P. (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics 12*, S2.
(141) Zerbino, D. R., and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research 18*, 821–829.
(142) Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu,

G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., and Wang, J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience 1*, 18–18.
(143) Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research 19*.
(144) Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010) Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol 28*, 503–510.
(145) Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol 28*, 516–520.
(146) Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G. K.-S., and Wang, J. (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* btu077.
(147) Hatem, A., Bozdag, D., Toland, A. E., and Catalyurek, U. V. (2013) Benchmarking short sequence mapping tools. *BMC Bioinformatics 14*, 184–184.
(148) Ben Langmead, Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology 10*, R25–R25.
(149) Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics 25*, 1754–1760.
(150) Deutsch, M., and Long, M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res 27*, 3219–3228.
(151) Misra, S., Narayanan, R., Lin, S., and Choudhary, A. (2010) FANGS: high speed sequence mapping for next generation sequencers. *SAC '10* 1539–1546.
(152) Novocraft. Novoalign. *www.novocraft.com*.
(153) Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E. E., and Sahinalp, S. C. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Publishing Group 7*, 576–577.
(154) Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet 41*, 1061–1067.
(155) Wu, T. D., and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics 26*, 873–881.
(156) Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics 9*, 128.
(157) Langmead, B. (2013) Introduction to the Burrows-Wheeler Transform and FM Index 1–12.
(158) Burrows, M., and Wheeler, D. J. (1994) A Block-sorting Lossless Data Compression Algorithm.
(159) Ferragina, P., and Manzini, G. (2000) Opportunistic data structures with applications. *SFCS-00* 390–398.
(160) Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics 25*, 1966–1967.
(161) Medina-Medina, N., Broka, A., Lacey, S., Lin, H., Klings, E. S., Baldwin, C. T., Steinberg, M. H., and Sebastiani, P. (2012) Comparing Bowtie and BWA to Align Short Reads from a RNA-Seq Experiment, in *6th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pp 197–207. Springer Berlin Heidelberg, Berlin, Heidelberg.
(162) Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics 25*, 2283–2285.
(163) Adams, J. (2008) Transcriptome: connecting the genome to gene function. Nature Education.
(164) Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R.

(2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics 12*, 231–231.
(165) You, N., Murillo, G., Su, X., Zeng, X., Xu, J., Ning, K., Zhang, S., Zhu, J., and Cui, X. (2012) SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics 28*, 643–650.
(166) Bertioli, D. J., Ozias-Akins, P., Chu, Y., and Dantas, K. M. (2014) The Use of SNP Markers for Linkage Mapping in Diploid and Tetraploid Peanuts. *G3: Genes| Genomes| …*.
(167) Kharabian-Masouleh, A., Waters, D. L. E., Reinke, R. F., and Henry, R. J. (2011) Discovery of polymorphisms in starch-related genes in rice germplasm by amplification of pooled DNA and deeply parallel sequencing†. *Plant Biotechnology Journal 9*, 1074–1085.
(168) Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2013) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol Ecol 23*, 40–69.
(169) Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research 22*, 568–576.
(170) Hamada, M., Wijaya, E., Frith, M. C., and Asai, K. (2011) Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. *Bioinformatics 27*, 3085–3092.
(171) Glaus, P., Honkela, A., and Rattray, M. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics 28*, 1721–1728.
(172) Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods 5*, 621–628.
(173) Natali, L., Cossu, R. M., Barghini, E., Giordani, T., Buti, M., Mascagni, F., Morgante, M., Gill, N., Kane, N. C., Rieseberg, L., and Cavallini, A. (2013) The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads. *BMC Genomics 14*, 1–1.
(174) Sveinsson, S., Gill, N., Kane, N. C., and Cronk, Q. (2013) Transposon fingerprinting using low coverage whole genome shotgun sequencing in Cacao (Theobroma cacao L.) and related species. *BMC Genomics 14*, 1–1.
(175) Hertweck, K. L. (2013) Assembly and comparative analysis of transposable elements from low coverage genomic sequence data in Asparagales. *Genome 56*, 487–494.
(176) Haas, B. J. TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies. *transposonpsi.sourceforge.net*.
(177) Wicker, T., Matthews, D. E., and Keller, B. (2002) TREP: a database for Triticeae repetitive elements. Trends in Plant Science.
(178) Rotmistrovsky, K., and Agarwala, R. (2011) BMTagger: Best Match Tagger for removing human reads from metagenomics datasets.
(179) Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal 17*, pp. 10–12.
(180) Kircher, M., Heyn, P., and Kelso, J. (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics 12*, 382.
(181) Canales, J., Bautista, R., Label, P., Gómez-Maldonado, J., Lesur, I., Fernández-Pozo, N., Rueda-López, M., Guerrero-Fernández, D., Castro-Rodríguez, V., Benzekri, H., Cañas, R. A., Guevara, M.-A., Rodrigues, A., Seoane, P., Teyssier, C., Morel, A., Ehrenmann, F., Le Provost, G., Lalanne, C., Noirot, C., Klopp, C., Reymond, I., García-Gutiérrez, A., Trontin, J.-F., Lelu-Walter, M.-A., Miguel, C., Cervera, M. T., Cantón, F. R., Plomion, C., Harvengt, L., Avila, C., Gonzalo Claros, M., and Cánovas, F. M. (2013) De novoassembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnology Journal 12*, 286–299.
(182) O'Neil, S. T., and Emrich, S. J. (2013) Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics 14*, 465–465.
(183) Fiehn, O. (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics 2*, 155–168.
(184) Booth, S. C., Weljie, A. M., and Turner, R. J. (2013) COMPUTATIONAL TOOLS FOR THE SECONDARY ANALYSIS OF METABOLOMICS EXPERIMENTS. *Computational and Structural Biotechnology Journal 4*, 1–13.
(185) Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis,

S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet 25*, 25–29.
(186) Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res 27*, 29–34.
(187) Sangwan, R. S., Tripathi, S., Singh, J., Narnoliya, L. K., and Sangwan, N. S. (2013) De novo sequencing and assembly of Centella asiatica leaf transcriptome for mapping of structural, functional and regulatory genes with special reference to secondary metabolism. *Gene 525*, 58–76.
(188) Rhee, S. Y., Wood, V., Dolinski, K., and Draghici, S. (2008) Use and misuse of the gene ontology annotations. *Nat Rev Genet 9*, 509–515.
(189) Primmer, C. R., Papakostas, S., Leder, E. H., Davis, M. J., and Ragan, M. A. (2013) Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol 22*, 3216–3241.
(190) Hill, D. P., Berardini, T. Z., Howe, D. G., and Van Auken, K. M. (2010) Representing ontogeny through ontology: a developmental biologist's guide to the gene ontology. *Mol. Reprod. Dev. 77*, 314–329.
(191) Barrett, A. J. (1995) Enzyme Nomenclature. Recommendations 1992. *European Journal of Biochemistry 232*, 1–1.
(192) Tipton, K., and Boyce, S. (2000) History of the enzyme nomenclature system. *Bioinformatics 16*, 34–40.
(193) Da Wei Huang, Sherman, B. T., and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res 37*, 1–13.
(194) Tipney, H., and Hunter, L. (2010) An introduction to effective use of enrichment analysis software. *Human Genomics 4*, 202–206.
(195) Da Wei Huang, Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology 8*, R183–R183.
(196) Sherman, B. T., Huang, D. W., Tan, Q., Guo, Y., Bour, S., Liu, D., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007) DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics 8*, 426.
(197) Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003) Identifying biological themes within lists of genes with EASE. *Genome biology 4*, R70.
(198) Xiao, Y., Di, P., Chen, J., Liu, Y., Chen, W., and Zhang, L. (2008) Characterization and expression profiling of 4-hydroxyphenylpyruvate dioxygenase gene (Smhppd) from Salvia miltiorrhiza hairy root cultures. *Mol Biol Rep 36*, 2019–2029.
(199) Jiang, Y., Xia, N., Li, X., Shen, W., Liang, L., Wang, C., Wang, R., Peng, F., and Xia, B. (2011) Molecular cloning and characterization of a phenylalanine ammonia-lyase gene (LrPAL) from Lycoris radiata. *Mol Biol Rep 38*, 1935–1940.
(200) Aza-González, C., Núñez-Palenius, H. G., and Ochoa-Alejo, N. (2010) Molecular biology of capsaicinoid biosynthesis in chili pepper (Capsicum spp.). *Plant Cell Rep 30*, 695–706.
(201) Petersen, M., Abdullah, Y., Benner, J., Eberle, D., Gehlen, K., Hücherig, S., Janiak, V., Kim, K. H., Sander, M., Weitzel, C., and Wolters, S. (2009) Evolution of rosmarinic acid biosynthesis. *Phytochemistry 70*, 1663–1679.
(202) Anarat-Cappillino, G., and Sattely, E. S. (2014) ScienceDirectThe chemical logic of plant natural product biosynthesis. *Current Opinion in Plant Biology 19*, 51–58.
(203) Luk, L., Bunn, S., Liscombe, D. K., and Facchini, P. J. (2007) Mechanistic studies on norcoclaurine synthase of benzylisoquinoline alkaloid biosynthesis: an enzymatic Pictet-Spengler reaction. *Biochemistry*.
(204) Samanani, N., Liscombe, D. K., and Facchini, P. J. (2004) Molecular cloning and characterization of norcoclaurine synthase, an enzyme catalyzing the first committed step in benzylisoquinoline alkaloid biosynthesis. *Plant J. 40*, 302–313.
(205) Gesell, A., Rolf, M., Ziegler, J., Chávez, M. L. D., Huang, F.-C., and Kutchan, T. M. (2009) CYP719B1 is salutaridine synthase, the C-C phenol-coupling enzyme of morphine biosynthesis in opium poppy. *J. Biol. Chem. 284*, 24432–24442.
(206) Stöckigt, J., Antonchick, A. P., Wu, F., and Waldmann, H. (2011) The Pictet-Spengler

Reaction in Nature and in Organic Chemistry. *Angew. Chem. Int. Ed. 50*, 8538–8564.
(207) Denisov, I. G., Makris, T. M., and Sligar, S. G. (2005) Structure and chemistry of cytochrome P450. *Chemical Reviews*.
(208) Zhao, Y.-J., Cheng, Q.-Q., Su, P., Chen, X., Wang, X.-J., Gao, W., and Huang, L.-Q. (2014) Research progress relating to the role of cytochrome P450 in the biosynthesis of terpenoids in medicinal plants. *Appl Microbiol Biotechnol 98*, 2371–2383.
(209) Frear, D. S., Swanson, H. R., and Tanaka, F. S. (1969) N-demethylation of substituted 3-(phenyl)-1-methylureas: Isolation and characterization of a microsomal mixed function oxidase from cotton. *Phytochemistry 8*, 2157–2169.
(210) Nomura, T., and Bishop, G. J. (2006) Cytochrome P450s in plant steroid hormone synthesis and metabolism. *Phytochem Rev 5*, 421–432.
(211) Mizutani, M., and Sato, F. (2011) Unusual P450 reactions in plant secondary metabolism. *Archives of Biochemistry and Biophysics 507*, 194–203.
(212) Guengerich, F. P., and Munro, A. W. (2013) Unusual Cytochrome P450 Enzymes and Reactions. *J. Biol. Chem. 288*, 17065–17073.
(213) Mizukami, H., Tabira, Y., and Ellis, B. E. (1993) Methyl jasmonate-induced rosmarinic acid biosynthesis in Lithospermum erythrorhizon cell suspension cultures. *Plant Cell Rep 12*, 706–709.
(214) Bjorklund, J. A., Frenzel, T., Rueffer, M., Kobayashi, M., Mocek, U., Fox, C., Beale, J. M., Groeger, S., Zenk, M. H., and Floss, H. G. (2002) Cryptic stereochemistry of berberine alkaloid biosynthesis. *J. Am. Chem. Soc. 117*, 1533–1545.
(215) Bauer, W., and Zenk, M. H. (1991) Two methylenedioxy bridge forming cytochrome P-450 dependent enzymes are involved in (< i> S</i>)-stylopine biosynthesis. *Phytochemistry*.
(216) Ikezawa, N., Iwasa, K., and Sato, F. (2009) CYP719A subfamily of cytochrome P450 oxygenases and isoquinoline alkaloid biosynthesis in Eschscholzia californica. *Plant Cell Rep*.
(217) Ikezawa, N., Tanaka, M., Nagayoshi, M., Shinkyo, R., Sakaki, T., Inouye, K., and Sato, F. (2003) Molecular Cloning and Characterization of CYP719, a Methylenedioxy Bridge-forming Enzyme That Belongs to a Novel P450 Family, from cultured Coptis japonica Cells. *J. Biol. Chem. 278*, 38557–38565.
(218) Ikezawa, N., Iwasa, K., and Sato, F. (2007) Molecular cloning and characterization of methylenedioxy bridge-forming enzymes involved in stylopine biosynthesis in Eschscholzia californica. *FEBS Journal*.
(219) Ikezawa, N., Iwasa, K., and Sato, F. (2008) Molecular cloning and characterization of CYP80G2, a cytochrome P450 that catalyzes an intramolecular C-C phenol coupling of (S)-reticuline in magnoflorine biosynthesis, from cultured Coptis japonica cells. *J. Biol. Chem. 283*, 8810–8821.
(220) Kraus, P. F., and Kutchan, T. M. (1995) Molecular cloning and heterologous expression of a cDNA encoding berbamunine synthase, a C--O phenol-coupling cytochrome P450 from the higher plant Berberis …, in.
(221) Pictet, A., and Spengler, T. (1911) Formation of isoquinoline derivatives by the action of methylal on phenylethylamine, phenylalanine and tyrosine. Ber. Dtsch. Chem. Ges.
(222) Stöckigt, J., and Zenk, M. H. (1977) Strictosidine (isovincoside): the key intermediate in the biosynthesis of monoterpenoid indole alkaloids. *Journal of the Chemical Society*.
(223) Stöckigt, J., and Zenk, M. H. (1977) Isovincoside (strictosidine), the key intermediate in the enzymatic formation of indole alkaloids. *FEBS letters*.
(224) Kutchan, T. M. (1993) Strictosidine: from alkaloid to enzyme to gene. *Phytochemistry*.
(225) Ma, X., Panjikar, S., Koepke, J., Loris, E., and Stöckigt, J. (2006) The structure of Rauvolfia serpentina strictosidine synthase is a novel six-bladed beta-propeller fold in plant proteins. *The Plant Cell 18*, 907–920.
(226) Facchini, P. J., and Deluca, V. (1995) EXPRESSION IN ESCHERICHIA-COLI AND PARTIAL CHARACTERIZATION OF 2 TYROSINE/DOPA DECARBOXYLASES FROM OPIUM POPPY. *Phytochemistry 38*.
(227) Liscombe, D. K., MacLeod, B. P., Loukanina, N., Nandi, O. I., and Facchini, P. J. (2005) Evidence for the monophyletic evolution of benzylisoquinoline alkaloid biosynthesis in angiosperms. *Phytochemistry 66*, 2501–2520.
(228) Ounaroon, A., Decker, G., Schmidt, J., Lottspeich, F., and Kutchan, T. M. (2003) (R,S)-Reticuline 7-O-methyltransferase and (R,S)-norcoclaurine 6-O-methyltransferase of Papaver somniferum – cDNA cloning and characterization of methyl transfer enzymes of

alkaloid biosynthesis in opium poppy. *The Plant Journal 36*, 808–819.
(229) Morishige, T., Tsujita, T., Yamada, Y., and Sato, F. (2000) Molecular Characterization of theS-Adenosyl-L-methionine:3‘‘-Hydroxy-N-methylcoclaurine 4-”O-Methyltransferase Involved in Isoquinoline Alkaloid Biosynthesis in Coptis japonica. *J. Biol. Chem. 275*, 23398–23405.
(230) Choi, K. B., Morishige, T., and Sato, F. (2001) Purification and characterization of coclaurine N-methyltransferase from cultured Coptis japonica cells. *Phytochemistry 56*, 649–655.
(231) Pauli, H. H., and Kutchan, T. M. (1998) Molecular cloning and functional heterologous expression of two alleles encoding (S)-N-methylcoclaurine 3'-hydroxylase (CYP80B1), a new methyl jasmonate-inducible cytochrome P-450-dependent mono-oxygenase of benzylisoquinoline alkaloid biosynthesis. *Plant J. 13*, 793–801.
(232) Frick, S., Kramell, R., and Kutchan, T. M. (2007) Metabolic engineering with a morphine biosynthetic P450 in opium poppy surpasses breeding. *Metab Eng 9*, 169–176.
(233) Pienkny, S., Brandt, W., Schmidt, J., Kramell, R., and Ziegler, J. (2009) Functional characterization of a novel benzylisoquinoline O-methyltransferase suggests its involvement in papaverine biosynthesis in opium poppy (Papaver somniferum L)
. *Plant J. 60*, 56–67.
(234) Ziegler, J., Voigtländer, S., Schmidt, J., Kramell, R., Miersch, O., Ammer, C., Gesell, A., and Kutchan, T. M. (2006) Comparative transcript and alkaloid profiling in Papaver species identifies a short chain dehydrogenase/reductase involved in morphine biosynthesis. *Plant J. 48*, 177–192.
(235) Lenz, R., and Zenk, M. H. (1995) Acetyl coenzyme A:salutaridinol-7-O-acetyltransferase from papaver somniferum plant cell cultures. The enzyme catalyzing the formation of thebaine in morphine biosynthesis. *J. Biol. Chem. 270*, 31091–31096.
(236) Grothe, T. (2001) Molecular Characterization of the Salutaridinol 7-O-Acetyltransferase Involved in Morphine Biosynthesis in Opium Poppy Papaver somniferum. *J. Biol. Chem. 276*, 30717–30723.
(237) Unterlinner, B., Lenz, R., and Kutchan, T. M. (1999) Molecular cloning and functional expression of codeinone reductase: the penultimate enzyme in morphine biosynthesis in the opium poppy Papaver somniferum. *The Plant Journal 18*, 465–475.
(238) Dittrich, H., and Kutchan, T. M. (1991) Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. *Proc Natl Acad Sci U S A 88*, 9969–9973.
(239) Kutchan, T. M., and Dittrich, H. (1995) Characterization and mechanism of the berberine bridge enzyme, a covalently flavinylated oxidase of benzophenanthridine alkaloid biosynthesis in plants. *J. Biol. Chem. 270*, 24475–24481.
(240) Facchini, P. J., Penzes, C., Johnson, A. G., and Bull, D. (1996) Molecular Characterization of Berberine Bridge Enzyme Genes from Opium Poppy. *PLANT PHYSIOLOGY 112*, 1669–1677.
(241) Winkler, A., Hartner, F., Kutchan, T. M., Glieder, A., and Macheroux, P. (2006) Biochemical evidence that berberine bridge enzyme belongs to a novel family of flavoproteins containing a bi-covalently attached FAD cofactor. *J. Biol. Chem. 281*, 21276–21285.
(242) Díaz Chávez, M. L., Rolf, M., Gesell, A., and Kutchan, T. M. (2011) Characterization of two methylenedioxy bridge-forming cytochrome P450-dependent enzymes of alkaloid formation in the Mexican prickly poppy Argemone mexicana. *Archives of Biochemistry and Biophysics 507*, 186–193.
(243) Liscombe, D. K., and Facchini, P. J. (2007) Molecular cloning and characterization of tetrahydroprotoberberine cis-N-methyltransferase, an enzyme involved in alkaloid biosynthesis in opium poppy. *J. Biol. Chem.*
(244) Takemura, T., Ikezawa, N., Iwasa, K., and Sato, F. (2013) Molecular cloning and characterization of a cytochrome P450 in sanguinarine biosynthesis from Eschscholzia californica cells. *Phytochemistry 91*, 100–108.
(245) Fujiwake, H., Suzuki, T., and Iwai, K. (1982) Intracellular distributions of enzymes and intermediates involved in biosynthesis of capsaicin and its analogues in Capsicum fruits. *Agricultural and Biological Chemistry*.
(246) Sukrasno, N., and Yeoman, M. M. (1993) Phenylpropanoid metabolism during growth and development of Capsicum frutescens fruits. *Phytochemistry 32*, 839–844.
(247) Mazourek, M., Pujar, A., Borovsky, Y., Paran, I., Mueller, L., and Jahn, M. M. (2009) A Dynamic Interface for Capsaicinoid Systems Biology. *PLANT PHYSIOLOGY 150*, 1806–1821.

(248) Stewart, C., Jr, Kang, B.-C., Liu, K., Mazourek, M., Moore, S. L., Yoo, E. Y., Kim, B.-D., Paran, I., and Jahn, M. M. (2005) The Pun1 gene for pungency in pepper encodes a putative acyltransferase. *The Plant Journal 42*, 675–688.

(249) Aluru, M. R., Mazourek, M., Landry, L. G., Curry, J., Jahn, M., and O'Connell, M. A. (2003) Differential expression of fatty acid synthase genes, Acl, Fat and Kas, in Capsicum fruit. *J Exp Bot 54*, 1655–1664.

(250) Lee, S. J., Suh, M.-C., Kim, S., Kwon, J.-K., Kim, M., Paek, K.-H., Choi, D., and Kim, B.-D. (2001) Molecular cloning of a novel pathogen-inducible cDNA encoding a putative acyl-CoA synthetase from Capsicum annuum L. *Plant Mol Biol 46*, 661–671.

(251) Kim, M., Kim, S., and Ki, B. D. (2001) Isolation of cDNA clones differentially accumulated in the placenta of pungent pepper by suppression subtractive hybridization. *Molecules and cells*.

(252) Lee, C. J., Yoo, E., Shin, J., Shin, J. H., and Lee, J. (2005) Non-pungent Capsicum contains a deletion in the capsaicinoid synthetase gene, which allows early detection of pungency with SCAR markers. *Molecules and …*.

(253) Huang, B., Yi, B., Duan, Y., Sun, L., Yu, X., Guo, J., and Chen, W. (2007) Characterization and expression profiling of tyrosine aminotransferase gene from Salvia miltiorrhiza (Dan-shen) in rosmarinic acid biosynthesis pathway. *Mol Biol Rep 35*, 601–612.

(254) Huang, B., Duan, Y., Yi, B., Sun, L., Lu, B., Yu, X., Sun, H., Zhang, H., and Chen, W. (2008) Characterization and expression profiling of cinnamate 4-hydroxylase gene from Salvia miltiorrhiza in rosmarinic acid biosynthesis pathway. *Russ J Plant Physiol 55*, 390–399.

(255) Xiao, Y., Gao, S., Di, P., Chen, J., Chen, W., and Zhang, L. (2009) Methyl jasmonate dramatically enhances the accumulation of phenolic acids in Salvia miltiorrhizahairy root cultures. *Physiologia Plantarum 137*, 1–9.

(256) Szabo, E., Thelen, A., and Petersen, M. (1999) Fungal elicitor preparations and methyl jasmonate enhance rosmarinic acid accumulation in suspension cultures of Coleus blumei. *Plant Cell Rep 18*, 485–489.

(257) Ivanov, I., Georgiev, V., and Pavlov, A. (2013) Elicitation of galanthamine biosynthesis by Leucojum aestivum liquid shoot cultures. *Journal of Plant Physiology 170*, 1122–1129.

(258) Tahchy, A. E., Boisbrun, M., Ptak, A., and Dupire, F. (2010) New method for the study of Amaryllidaceae alkaloid biosynthesis using biotransformation of deuterium-labeled precursor in tissue cultures. Acta Biochimica ….

(259) Sato, S., Nakamura, Y., Kaneko, T., Katoh, T., Asamizu, E., Kotani, H., and Tabata, S. (2000) Structural analysis of Arabidopsis thaliana chromosome 5. X. Sequence features of the regions of 3,076,755 bp covered by sixty P1 and TAC clones. *DNA Res. 7*, 31–63.

(260) Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., Pham, P., Cheuk, R., Karlin-Newmann, G., Liu, S. X., Lam, B., Sakano, H., Wu, T., Yu, G., Miranda, M., Quach, H. L., Tripp, M., Chang, C. H., Lee, J. M., Toriumi, M., Chan, M. M. H., Tang, C. C., Onodera, C. S., Deng, J. M., Akiyama, K., Ansari, Y., Arakawa, T., Banh, J., Banno, F., Bowser, L., Brooks, S., Carninci, P., Chao, Q., Choy, N., Enju, A., Goldsmith, A. D., Gurjal, M., Hansen, N. F., Hayashizaki, Y., Johnson-Hopson, C., Hsuan, V. W., Iida, K., Karnes, M., Khan, S., Koesema, E., Ishida, J., Jiang, P. X., Jones, T., Kawai, J., Kamiya, A., Meyers, C., Nakajima, M., Narusaka, M., Seki, M., Sakurai, T., Satou, M., Tamse, R., Vaysberg, M., Wallender, E. K., Wong, C., Yamamura, Y., Yuan, S., Shinozaki, K., Davis, R. W., Theologis, A., and Ecker, J. R. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science 302*, 842–846.

(261) Prabhu, P. R., and Hudson, A. O. (2010) Identification and partial characterization of an L-tyrosine aminotransferase (TAT) from Arabidopsis thaliana. *Biochemistry research …*.

(262) Facchini, P. J., Penzes-Yost, C., Samanani, N., and Kowalchuk, B. (1998) Expression patterns conferred by tyrosine/dihydroxyphenylalanine decarboxylase promoters from opium poppy are conserved in transgenic tobacco. *PLANT PHYSIOLOGY 118*, 69–81.

(263) Facchini, P. J., and Deluca, V. (1994) DIFFERENTIAL AND TISSUE-SPECIFIC EXPRESSION OF A GENE FAMILY FOR TYROSINE DOPA DECARBOXYLASE IN OPIUM POPPY. *J. Biol. Chem. 269*.

(264) Maldonado-Mendoza, I. (1996) Molecular analysis of a new member of the opium poppy tyrosine/3,4- dihydroxyphenylalanine decarboxylase gene family. *PLANT PHYSIOLOGY 110*, 43–49.

(265) Bevan, M., Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Müller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., Schmidtheini, T., Reichert, B., Portatelle, D., Perez-

Alonso, M., Boutry, M., Bancroft, I., Vos, P., Hoheisel, J., Zimmermann, W., Wedler, H., Ridley, P., Langham, S. A., McCullagh, B., Bilham, L., Robben, J., Van der Schueren, J., Grymonprez, B., Chuang, Y. J., Vandenbussche, F., Braeken, M., Weltjens, I., Voet, M., Bastiaens, I., Aert, R., Defoor, E., Weitzenegger, T., Bothe, G., Ramsperger, U., Hilbert, H., Braun, M., Holzer, E., Brandt, A., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Lankhorst, R. K., Rose, M., Hauf, J., Kötter, P., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Van den Daele, H., De Keyser, A., Buysshaert, C., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Rogers, J., Cronin, A., Quail, M., Bray-Allen, S., Clark, L., Doggett, J., Hall, S., Kay, M., Lennard, N., McLay, K., Mayes, R., Pettett, A., Rajandream, M. A., Lyne, M., Benes, V., Rechmann, S., Borkova, D., Blöcker, H., Scharfe, M., Grimm, M., Löhnert, T. H., Dose, S., de Haan, M., Maarse, A., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fartmann, B., Granderath, K., Dauner, D., Herzl, A., Neumann, S., Argiriou, A., Vitale, D., Liguori, R., Piravandi, E., Massenet, O., Quigley, F., Clabauld, G., Mündlein, A., Felber, R., Schnabl, S., Hiller, R., Schmidt, W., Lecharny, A., Aubourg, S., Chefdor, F., Cooke, R., Berger, C., Montfort, A., Casacuberta, E., Gibbons, T., Weber, N., Vandenbol, M., Bargues, M., Terol, J., Torres, A., Perez-Perez, A., Purnelle, B., Bent, E., Johnson, S., Tacon, D., Jesse, T., Heijnen, L., Schwarz, S., Scholler, P., Heber, S., Francs, P., Bielke, C., Frishman, D., Haase, D., Lemcke, K., Mewes, H. W., Stocker, S., Zaccaria, P., Wilson, R. K., la Bastide, de, M., Habermann, K., Parnell, L., Dedhia, N., Gnoj, L., Schutz, K., Huang, E., Spiegel, L., Sehkon, M., Murray, J., Sheet, P., Cordes, M., Abu-Threideh, J., Stoneking, T., Kalicki, J., Graves, T., Harmon, G., Edwards, J., Latreille, P., Courtney, L., Cloud, J., Abbott, A., Scott, K., Johnson, D., Minx, P., Bentley, D., Fulton, B., Miller, N., Greco, T., Kemp, K., Kramer, J., Fulton, L., Mardis, E., Dante, M., Pepin, K., Hillier, L., Nelson, J., Spieth, J., Ryan, E., Andrews, S., Geisel, C., Layman, D., Du, H., Ali, J., Berghoff, A., Jones, K., Drone, K., Cotton, M., Joshu, C., Antonoiu, B., Zidanic, M., Strong, C., Sun, H., Lamar, B., Yordan, C., Ma, P., Zhong, J., Preston, R., Vil, D., Shekher, M., Matero, A., Shah, R., Swaby, I., O'Shaughnessy, A., Rodriguez, M., Hoffman, J., Till, S., Granat, S., Shohdy, N., Hasegawa, A., Hameed, A., Lodhi, M., Johnson, A., Chen, E., Marra, M., Martienssen, R., and McCombie, W. R. (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. *Nature 402*, 769–777.
(266) Liscombe, D. K., MacLeod, B. P., Loukanina, N., Nandi, O. I., and Facchini, P. J. (2005) Evidence for the monophyletic evolution of benzylisoquinoline alkaloid biosynthesis in angiosperms (vol 66, pg 1374, 2005). *Phytochemistry 66*, 2500–2520.
(267) Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J., Ronning, C. M., Koo, H. L., Moffat, K. S., Cronin, L. A., Shen, M., Pai, G., Van Aken, S., Umayam, L., Tallon, L. J., Gill, J. E., Adams, M. D., Carrera, A. J., Creasy, T. H., Goodman, H. M., Somerville, C. R., Copenhaver, G. P., Preuss, D., Nierman, W. C., White, O., Eisen, J. A., Salzberg, S. L., Fraser, C. M., and Venter, J. C. (1999) Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. *Nature 402*, 761–768.
(268) Minami, H., Dubouzet, E., Iwasa, K., and Sato, F. (2007) Functional analysis of norcoclaurine synthase in Coptis japonica. *J. Biol. Chem. 282*, 6274–6282.
(269) Berkner, H., Schweimer, K., Matecko, I., and Rösch, P. (2008) Conformation, catalytic site, and enzymatic mechanism of the PR10 allergen-related enzyme norcoclaurine synthase. *Biochemical Journal 413*, 281–290.
(270) Berkner, H., Engelhorn, J., Liscombe, D. K., Schweimer, K., Wöhrl, B. M., Facchini, P. J., Rösch, P., and Matecko, I. (2007) High-yield expression and purification of isotopically labeled norcoclaurine synthase, a Bet v 1-homologous enzyme, from Thalictrum flavum for NMR studies. *Protein Expr. Purif. 56*, 197–204.
(271) Sato, F., Tsujita, T., Katagiri, Y., Yoshida, S., and Yamada, Y. (1994) Purification and characterization of S-adenosyl-L-methionine: norcoclaurine 6-O-methyltransferase from cultured Coptis japonica cells. *Eur. J. Biochem. 225*, 125–131.
(272) Choi, K. B., Morishige, T., Shitan, N., Yazaki, K., and Sato, F. (2002) Molecular Cloning and Characterization of CoclaurineN-Methyltransferase from Cultured Cells of Coptis japonica. *J. Biol. Chem. 277*, 830–835.
(273) Samanani, N., Park, S.-U., and Facchini, P. J. (2005) Cell type-specific localization of transcripts encoding nine consecutive enzymes involved in protoberberine alkaloid biosynthesis. *The Plant Cell 17*, 915–926.
(274) Kato, N., Dubouzet, E., Kokabu, Y., Yoshida, S., Taniguchi, Y., Dubouzet, J. G., Yazaki, K., and Sato, F. (2007) Identification of a WRKY protein as a transcriptional regulator of benzylisoquinoline alkaloid biosynthesis in Coptis japonica. *Plant Cell Physiol 48*, 8–18.

(275) Alcantara, J., Bird, D. A., Franceschi, V. R., and Facchini, P. J. (2005) Sanguinarine biosynthesis is associated with the endoplasmic reticulum in cultured opium poppy cells after elicitor treatment. *PLANT PHYSIOLOGY 138*, 173–183.
(276) Lee, E. J., Hwang, I. K., Kim, N. Y., Lee, K. L., Han, M. S., Lee, Y. H., Kim, M. Y., and Yang, M. S. (2010) An assessment of the utility of universal and specific genetic markers for opium poppy identification. *J. Forensic Sci. 55*, 1202–1208.
(277) Morishige, T., Tamakoshi, M., Takemura, T., and Sato, F. (2010) Molecular characterization of O-methyltransferases involved in isoquinoline alkaloid biosynthesis in Coptis japonica. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences 86*.
(278) Inui, T., Tamura, K.-I., Fujii, N., Morishige, T., and Sato, F. (2007) Overexpression of Coptis japonica norcoclaurine 6-O-methyltransferase overcomes the rate-limiting step in Benzylisoquinoline alkaloid biosynthesis in cultured Eschscholzia californica. *Plant Cell Physiol 48*, 252–262.
(279) Kang, S., Kang, K., Chung, G. C., Choi, D., Ishihara, A., Lee, D.-S., and Back, K. (2006) Functional analysis of the amine substrate specificity domain of pepper tyramine and serotonin N-hydroxycinnamoyltransferases. *PLANT PHYSIOLOGY 140*, 704–715.
(280) Farmer, M. J., Czernic, P., Michael, A., and Negrel, J. (1999) Identification and characterization of cDNA clones encoding hydroxycinnamoyl-CoA:tyramine N-hydroxycinnamoyltransferase from tobacco. *Eur. J. Biochem. 263*, 686–694.
(281) Budiman, M. A., Mao, L., Wood, T. C., and Wing, R. A. (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Research 10*, 129–136.
(282) Wanner, L. A., Li, G., Ware, D., Somssich, I. E., and Davis, K. R. (1995) The phenylalanine ammonia-lyase gene family in Arabidopsis thaliana. *Plant Mol Biol 27*, 327–338.
(283) Kim, D. S., and Hwang, B. K. (2014) An important role of the pepper phenylalanine ammonia-lyase gene (PAL1) in salicylic acid-dependent signalling of the defence response to microbial pathogens. *J Exp Bot 65*, 2295–2306.
(284) Kim, K. H., Janiak, V., and Petersen, M. (2004) Purification, cloning and functional expression of hydroxyphenylpyruvate reductase involved in rosmarinic acid biosynthesis in cell cultures of Coleus blumei. *Plant Mol Biol 54*, 311–323.
(285) Ramos-Onsins, S. E., Puerma, E., Balañá-Alcaide, D., Salguero, D., and Aguadé, M. (2008) Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in Arabidopsis thaliana. *Mol Ecol 17*, 1211–1223.
(286) Ehlting, J., Büttner, D., Wang, Q., Douglas, C. J., Somssich, I. E., and Kombrink, E. (1999) Three 4-coumarate:coenzyme A ligases in Arabidopsis thaliana represent two evolutionarily divergent classes in angiosperms. *Plant J. 19*, 9–20.
(287) Aza-González, C., and Herrera-Isidrón, L. (2013) Anthocyanin accumulation and expression analysis of biosynthesis-related genes during chili pepper fruit development. *Biologia Plantarum*.
(288) Feinbaum, R. L., and Ausubel, F. M. (1988) Transcriptional regulation of the Arabidopsis thaliana chalcone synthase gene. *Molecular and Cellular Biology*.
(289) Nguyen, P., and Cin, V. D. (2009) The role of light on foliage colour development in coleus (Solenostemon scutellarioides(L.) Codd). *Plant Physiol Biochem 47*, 934–945.
(290) WU, S., and Guo, J. (2012) Cloning and Analysis of Caffeic Acid O-methyltransferase Gene (SmCOMT1) from Salvia miltiorrhiza Bge. *Bulletin of Botanical ….*
(291) Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc 4*, 44–57.
(292) Piatigorsky, J. (2001) Dual use of the transcriptional repressor (CtBP2)/ribbon synapse (RIBEYE) gene: how prevalent are multifunctional genes? *Trends in Neurosciences 24*, 555–557.
(293) Folkers, U., Kirik, V., Schöbinger, U., Falk, S., Krishnakumar, S., Pollock, M. A., Oppenheimer, D. G., Day, I., Reddy, A. S. M., Jürgens, G., Hülskamp, M., and Reddy, A. R. (2002) The cell morphogenesis gene ANGUSTIFOLIA encodes a CtBP/BARS-like protein and is involved in the control of the microtubule cytoskeleton. *EMBO J. 21*, 1280–1288.
(294) Schmitz, F., Königstorfer, A., and Südhof, T. C. (2000) RIBEYE, a Component of Synaptic Ribbons. *Neuron 28*, 857–872.
(295) Nelson, D., and Werck-Reichhart, D. (2011) A P450-centric view of plant evolution. *The Plant Journal 66*, 194–211.
(296) Kamenetsky, R., and Okubo, H. (2012) Ornamental Geophytes. CRC Press.

(297) Raboin, L.-M., Carreel, F., Noyer, J.-L., Baurens, F.-C., Horry, J.-P., Bakry, F., Montcel, Du, H. T., Ganry, J., Lanaud, C., and Lagoda, P. J. L. (2005) Diploid Ancestors of Triploid Export Banana Cultivars: Molecular Identification of 2n Restitution Gamete Donors and n Gamete Donors. *Mol Breeding 16*, 333–341.
(298) BRANDHAM, P. (1999) New chromosome counts in Narcissus cultivars. *RHS Daffodil & Tulip Yearbook*.
(299) Hegarty, M. J., and Hiscock, S. J. (2008) Genomic clues to the evolutionary success of review polyploid plants. *Curr. Biol. 18*, R435–R444.
(300) Pignatta, D., and Comai, L. (2009) Parental squabbles and genome expression: lessons from the polyploids. *J. Biol. 8*, 43.
(301) Rambani, A., Page, J. T., and Udall, J. A. (2014) Polyploidy and the petal transcriptome of Gossypium. *BMC Plant Biol 14*.
(302) Hegarty, M. (2011) Hybridization: Expressing Yourself in a Crowd. *Curr. Biol. 21*, R254–R255.
(303) Auer, P. L., and Doerge, R. W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics 185*, 405–416.
(304) Paddon, C. J., and Keasling, J. D. (2014) Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol. 12*, 355–367.
(305) Nakagawa, A., Minami, H., Kim, J.-S., Koyanagi, T., Katayama, T., Sato, F., and Kumagai, H. (2011) A bacterial platform for fermentative production of plant alkaloids. *Nature Communications 2*, 326–8.
(306) Verma, A., Laakso, I., Seppänen-Laakso, T., Huhtikangas, A., and Riekkola, M.-L. (2007) A Simplified Procedure for Indole Alkaloid Extraction from Catharanthus roseus Combined with a Semi-synthetic Production Process for Vinblastine 1–9.
(307) D'haeseleer, P. (2006) What are DNA sequence motifs? *Nat Biotechnol 24*, 423–425.
(308) Wang, Q., Yuan, F., Pan, Q., Li, M., Wang, G., Zhao, J., and Tang, K. (2009) Isolation and functional analysis of the Catharanthus roseus deacetylvindoline-4-O-acetyltransferase gene promoter. *Plant Cell Rep 29*, 185–192.
(309) Ouwerkerk, P. B., and Memelink, J. (1999) A G-box element from the Catharanthus roseus strictosidine synthase (Str) gene promoter confers seed-specific expression in transgenic tobacco plants. *Mol. Gen. Genet. 261*, 635–643.