Investigating lipid and secondary metabolisms in plants by next-generation sequencing

JIN JINGJING

NATIONAL UNIVERSITY OF SINGAPORE

2014

Investigating lipid and secondary metabolisms in plants by next-generation sequencing

JIN JINGJING

(B.COMP., SCU)
(B.ECOM., SCU)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has not been submitted for any degree in any university previously.

Jin Jingjing

11th June 2014

Acknowledgements

First and foremost, I thank my supervisor Professor Limsoon Wong, for investing a huge amount of time in advising my doctoral work. Over the past years, I have benefited from his excellent guidance and persistent support. Working with him has been pleasant for me. I have learnt a lot from him in many aspects of doing research.

I also thank Professor Nam-Hai Chua, a leading plant scientist and my second mentor. During many discussions with him, I have learnt a lot of biology and attitude to research from him.

I am grateful to several principal investigators in Temasek Life Sciences Laboratory---in particular, Dr Jian Ye, Dr GenHua Yue, Dr Rajani Sarojam and Dr In-Cheol Jang---for their useful suggestions, sharing and discussion with me. I appreciate also a gift from Temasek Life Science Laboratory that supported the fifth year of my PhD studies.

I thank my parents Jin, Ting and Bai, Caiqin for their support and encouragement, which greatly motivate me to fully concentrate on my research.

I thank my seniors Dr Difeng Dong, Dr Guimei Liu, Dr Wilson Wen Bin Goh, Dr Jun Liu, Dr Huan Wang, Dr Shulin Deng and Dr Huiwen Wu, for teaching me so much about bioinformatics and plant biology, when I was a fresh PhD student.

Finally, I appreciate the friendship and support of my friends: Yong Lin, Mo Chen, Pingzhi Zhao, Hufeng Zhou, Haojun Zhang and many others. I want to express

my sincerest gratitude to them for the collaborative and useful discussions.

Contents

Summary vi
List of Tables viii
List of Figures x
1 Introduction 1
1.1 Motivation 2
1.1.1 Lipid 2
1.1.2 Secondary metabolism 4
1.1.3 Research challenges 5
1.2 Thesis contribution 6
1.3 Thesis organization 7
1.4 Declaration 8
2 Related work 9
2.1 Next-generation sequencing 9
2.2 Whole-genome sequencing 12
2.3 Genome resequencing 16
2.4 Molecular marker development 18
2.5 Transcriptome sequencing 19
2.6 Non-coding RNA characterization 21
3 reference-based genome assembly 25
3.1 Background 26
3.1.1 OLC-based assembly methods 26
3.1.2 DBG-based assembly methods 28
3.1.3 Reference-based genome assembly 29
3.2 Methods 30
3.2.2 Mis-assembled scaffold identification and correction 33
3.2.3 Alignment to reference genome 36
3.2.4 Repeat scaffold identification 36
3.2.5 Overlap scaffold identification 38
3.3 Results 39
3.3.1 Evaluation on gold-standard dataset 39
3.3.2 Evaluation of mis-assembly detection component 40
3.3.3 Evaluation of repeat-scaffold detection component 43
3.3.4 Evaluation of overlap-scaffold detection component 46
3.3.5 Comparison between de-novo and reference-based genome assembly 47
3.4 Conclusions 49
4 Application on oil palm 50
4.1 Background 51
4.2 Methods 53
4.2.1 Whole-genome short-gun (WGS) sequencing for oil palm 53
4.2.2 Reference-based genome assembly 54

4.3 Results 54
4.3.1 Evaluation method 54
4.3.2 Comparison between de novo assembly and reference-based assembly 55
4.3.3 Comparison between ABACAS and our proposed method 57
4.3.3.1 Effect of mis-assembly identification component 57
4.3.3.2 Effect of the repeat-scaffold identification component 58
4.4 Evaluation of Dura draft genome 60
4.4.1 EST coverage 60
4.4.2 Completeness of draft genome 61
4.4.3 Linkage map 62
4.5 Annotation of Dura draft genome 63
4.5.1 Repeat annotation 64
4.5.1.1 De novo identification of repeat sequence 64
4.5.1.2 Identification of known TEs 64
4.5.1.3 Tandem repeats 65
4.5.2 Gene annotation 66
4.5.2.1 De novo gene prediction 66
4.5.2.2 Evidence-based gene prediction 66
4.5.2.3 Reference gene set 69
4.5.2.4 Gene Function Annotation 69
4.5.3 NcRNA annotation 71
4.5.3.1 Identification of tRNAs 72
4.5.3.2 Identification of rRNAs 73
4.5.3.3 Identification of other small ncRNAs 74
4.5.3.4 Identification of long intergenic noncoding RNA (lincRNA) 75
4.6 Gene family for fatty acid pathway 79
4.7 Homologous genes 80
4.8 Whole-genome duplication 81
4.9 Evolution history of oil palm 84
4.9.1 Overview of diversity for oil palm 86
4.9.2 Structure and population analysis for oil palm 88
4.10 Conclusion 93
5 Visualization of various genome information 95
5.1 An online database to deposit, browse and download genome element 95
5.2 Visualizing detail information for transcript unit 96
5.3 Visualizing relative expression level across the whole genome 97
5.4 Visualizing smRNA abundance across the whole genome 98
5.5 BLAST tool 99
5.6 Conclusions 100
6 Weighted pathway approach 101
6.1 Background 104
6.1.1 Co-regulated genes 106
6.1.2 Over-representation analysis (ORA) 106
6.1.3 Direct-group Analysis 107
6.1.4 Network-based Analysis 108
6.1.5 Model-based Analysis 109
6.2 Methods 109

6.2.1 Preparatory step 1: Database of plant metabolic pathway 111
6.2.2 Preparatory step 2: Calculation of enzyme gene expression level 112
6.2.3 Main step 1: Relative gene expression level of enzyme 113
6.2.4 Main step 2: Identifying significant pathways 117
6.2.5 Main step 3: Extracting sub-networks 119
6.3 Results 120
6.3.1 Plant metabolic pathway database 120
6.3.2 Validity of weighted pathway approach 122
6.3.2.1 *VTE2* mutant 122
6.3.2.2 *SID2* mutant 126
6.4 Conclusion 131
7 Application on secondary metabolisms 133
7.1 Background 133
7.2 Methods 136
7.2.1 RNA sequencing 136
7.2.2 Weighted pathway analysis 138
7.3 Results 139
7.3.1 Results for RNA-seq 139
7.3.2 Results for weighted pathway approach 142
7.3.2.1 Enriched pathway for weighted pathway approach 142
7.3.2.2 Comparison between GC-MS result and weighted pathway approach result 143
7.3.2.3 Comparison with other pathway analysis methods 144
7.3.2.4 Comparison between results based on absolute expression level and relative expression level 146
7.3.2.5 Comparison between results based on transcriptome analysis and weighted pathway approach 148
7.4 Conclusion 152
8 Conclusion 153
8.1 Summary 153
8.2 Future work 156
BIBLIOGRAPHY 157

SUMMARY

Plant metabolites are compounds synthesized by plants for essential functions, such as growth and development (primary metabolites, such as lipid), and specific functions, such as pollinator attraction and defense against herbivores (secondary metabolites). Many of them are still used directly, or as derivatives, to treat a wide range of diseases for humans. There is a demand to explore the biosynthesis of different plant metabolites and improve their yield.

Next-generation sequencing (NGS) techniques have been proved valuable in the investigation of different plant metabolisms. However, genome resources for primary metabolites, especially lipids, are very scarce. Similarly, using NGS, most current studies of secondary metabolites just focus on known function/metabolic pathways. Hence, in this dissertation, we systemically investigate plant lipid metabolisms and secondary metabolisms by several different studies.

We first develop a reference-based genome assembly pipeline, including mis-assembled scaffold and repeat scaffold identification components. From the evaluation on a gold-standard dataset, we find that these major components in our pipeline have relatively high accuracy.

Next, we use our proposed reference-based genome assembly pipeline to construct a draft genome for Dura oil palm. Then, annotations---including protein-coding genes, small noncoding RNAs and long noncoding RNAs---are done for the draft genome. In addition, by resequencing 12 different oil palm strains, around 21 million high-quality single-nucleotide polymorphisms (SNPs) are found. Using these population SNP data, lots of sites with a high level of sequence diversity among different oil palms are identified. Some of these variants are associated with important biological functions, which can guide future breeding efforts for oil palm.

At the same time, a GBrowse-based database with a BLAST tool is developed to visualize different genome information of oil palm. It provides location information, expression information and structure information for different elements, such as protein-coding genes and noncoding RNAs.

In order to predict new functions/metabolisms for plants, a weighted pathway approach is proposed, which tries to consider dependencies between different pathways. From the validation results on two different models, we find that the weighted pathway approach is much more reasonable than traditional pathway analysis methods which do not take into consideration dependencies across pathways.

After applying this weighted pathway approach to an RNA-seq dataset from spearmint, several new functions and metabolisms are uncovered, such as energy-related functions, sesquiterpene and diterpene synthesis. The presence of most of these new metabolites is consistent with GC-MS results, and mRNAs encoding related enzymes have also been verified by q-PCR experiment.

LIST OF TABLES

Table 1.1 Oil production per weight for oil crops [Wikipedia] 3

Table 2.1 Comparison of performance and advantages of various NGS platform [27]... 10

Table 3.1 Comparison between different assemblers on short reads example for a known genome [90] 27
Table 3.2 Comparison of running time (Runtime) and RAM for different de novo assembly method [100]. SE denotes single-end sequencing dataset. PE denotes pair-end sequencing dataset. E.coli, C.ele, H.sap-2, H.sap-3 denotes four different test dataset. Second column denotes different de novo assembly method. ---denotes RAM of the server is not enough or running time too long (>10 days). s denotes second. MB denotes megabytes. 33
Table 3.3 Statistic of sequencing information for gold dataset 40
Table 3.4 Mis-assembly result based on the gold-standard data from Assemblathon 1 [103]. The number means the average number of mis-assembled scaffolds reported by our method.. 42
Table 3.5 Repeat scaffold result based on the gold-standard data from Assemblathon 1 [103]. The number is the average number of scaffolds mapped to multiple locations in the reference genome for different methods. 44
Table 3.6 Average number of overlap scaffold groups based on the gold-standard data from Assemblathon 1 [103] at different coverage. 47

Table 4.1 Sequence library for Dura by next-generation sequencing platform............... 54
Table 4.2 Comparison between different de novo assembly tools on Contig level 56
Table 4.3 Comparison between de novo assembly methods and our proposed reference-based method.. 56
Table 4.4 Comparison between ABACAS and our method.. 57
Table 4.5 Mis-assembly information in our pipeline 58
Table 4.6 Statistic for the repeat scaffolds 58
Table 4.7 Statistic result for the EST coverage of the Dura draft genome 61
Table 4.8 Repeat statistics for oil palm draft genome 65
Table 4.9 Comparison of oil palm with other plants on gene number, average exon/intron length and other parameters. Gene density: the number of gene per 10kb 69
Table 4.10 Compare oil palm with other plants on different class of tRNAs.................. 72
Table 4.11 Overview information of ncRNAs on oil palm draft genome 73
Table 4.12 Statistic information for the gene, lincRNA and miRNA identified by RNA seq data set 78
Table 4.13 The number of genes in fatty acid biosynthesis pathways for each plants.... 80
Table 4.14 Description of 12 oil palm strains 85
Table 4.15 SNP number between each oil palm strains and reference genome............. 86

Table 6.1 Statistic information for different pathway database 120
Table 6.2 Expression level for enzyme EC-1.13.11.27. WT and VTE2: denote expression level using absolute expression level; WT_weighted and VTE2_weighted: denote

using our weighted pathway model 123
Table 6.3 Mean value for different pathway WT and VTE2 denotes mean value using absolute expression level; WT_weighted and VTE2_weighted denotes the mean value using our weighted pathway model 124
Table 6.4 Rank for different pathways based on relative expression level for VTE2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than the mean in the wild type or mutant; rank (mean & size>3) denotes rank using pathways having relative expression level more than mean in wild type or mutant and size should be more than 3; rank (sub-network) denotes rank using sub-networks. 125
Table 6.5 Rank for different pathways based on absolute expression level for VTE2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than the mean in the wild type or mutant; rank (sub-network) denotes rank using sub-networks. 125
Table 6.6 Expression level for enzyme EC-4.2.3.5 in WT and ICS mutant. WT and Mutant denote the absolute expression level. WT_weighted and Mutant_weighted denote the relative expression level by our weighted pathway model............................ 128
Table 6.7 Mean value for different pathway. WT and ICS denotes mean value using absolute expression. WT_weighted and ICS_weighted denote mean value using relative expression. 129
Table 6.8 Rank for different pathways based on relative expression level for SID2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than mean in WT or mutant; rank (mean & size>3 130
Table 6.9 Rank for different pathways based on absolute expression level for SID2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than mean in WT or mutant; rank (sub-network).. 131

Table 7.1 Statistic for RNA seq results 136
Table 7.2 Assembly results for the plant samples in our study 139
Table 7.3 Top 20 enrichment pathway for trichome and other tissue in mint by our weighted pathway model Where each row denotes a pathway; column (leaf, root, leaf-trichome, trichome) denotes the overall expression level for a pathway by mean value of the enzyme in the pathway; FC denotes fold change between trichome and leaf-trichome using mean overall value; median and sum denotes overall expression level for trichome tissue by median value and sum value of the enzymes in the pathway; Pearson denotes the score for a pathway by the average Pearson correlation among one pathway; scorePAGE denote the score computed by scorePAGE method [183] 143
Table 7.4 Top 20 enriched pathway for mint by absolute expression level for each enzyme. Trichome denotes the overall expression level using the absolute value; our method denotes overall expression level for trichome tissue based on our solution, rank is the rank for each pathway in our solution; hub compound and hub enzyme is the number for hub compound and enzyme. 147

LIST OF FIGURES

Figure 3.1 Pipeline of our proposed reference-based genome assembly pipeline 31
Figure 3.2 An example of a mis-assembled scaffold [scaffold148]. a. the coverage across the scaffold 148 by insert size of pair end reads b. the detail alignment information for scaffold 148 after aligning to the reference genome. In this figure, t denotes target reference genome, q denotes query assembly scaffolds. 34
Figure 3.3 Model of assembly by pair end reads. The arrow denotes pair end reads 35
Figure 3.4 An example coverage comparison between a repeat scaffold and a non-repeat scaffold 38
Figure 3.5 A method to deal with the overlap scaffolds 39
Figure 3.6 Average number of assembled scaffolds by different de novo assembly methods 42
Figure 3.7 Percentage of correct mis-assembled scaffolds reported by our method for each de novo assembly method under different coverage of the raw genome 43
Figure 3.8 Recall for our repeat scaffold identification component 44
Figure 3.9 Precision for our repeat scaffold identification component 46
Figure 3.10 N50 for different method under different coverage of genome. 48
Figure 3.11 Final genome coverage by de novo assembly methods. Genome coverage=total number of bases of final scaffolds/genome size 49

Figure 4.1 Trends in global production of major plant oils [1] 50
Figure 4.2 Plant genomes which have been finished [111] 53
Figure 4.3 Pie chart of the increased scaffold located in reference genome, comparing to ABACAS 59
Figure 4.4 Relationship between linkage map and scaffolds in the draft genome of oil palm 63
Figure 4.5 An overview of the gene prediction results by MAKER2 [126], visualized based on our developed database [137] 68
Figure 4.6 The number of homologous genes in each species 71
Figure 4.7 Pipeline for identification of long intergenic noncoding RNA 76
Figure 4.8 Expression level of protein coding gene, pre-miRNA and lincRNA 79
Figure 4.9 Venn graph of homologs between oil palm, date palm, *Vitis* and rice 81
Figure 4.10 a: synteny region between oil palm and soybean b: synteny region between oil palm and *Vitis* 82
Figure 4.11 Detail synteny regions for one chromosome from oil palm 83
Figure 4.12 The synteny region in the detail location of each chromosome. a Synteny region between oil palm and date palm b Synteny region between soybean and oil palm c Synteny region between oil palm and *Vitis* 83
Figure 4.13 Statistic for different SNP categories of oil palm 87
Figure 4.14 Population genetic analysis of oil palm a: neighbor-joining tree for 12 different oil palm strains b: PCA result for 12 different oil palm strains c: Bayesian clustering (STRUCTURE, K=3) d:iHS score for different diversity sites across all chromosomes 89

Figure 4.15 Enriched GO terms for high-diversity gene locus Orange: biological process Green: cellular component Blue: Molecular function 91
Figure 4.16 Enriched GO terms for low-diversity gene locus Orange: biological process Green: cellular component Blue: Molecular function 92
Figure 4.17 Global overview about chromosome of oil palm a: chromosome information b: iHS score distribution c: gene density d: repeat density e: segmental duplication in genome 93

Figure 5.1 Snapshot of the GBrowse database to visualize the genome element 96
Figure 5.2 An example of detail information for transcript unit in the database........... 97
Figure 5.3 Snapshot for the expression level of our database 98
Figure 5.4 Snapshot of the BLAST function for oil palm database 99

Figure 6.1 Simplified schematic overview of the biosynthesis of the main secondary metabolites stored and/or secreted by glandular trichome cells. Major pathway names are shown in red, key enzymes or enzyme complexes in purple, and stored and/or secreted compounds in blue. [168] 101
Figure 6.2 Glandular trichomes in section Lycopersicon. [168] 103
Figure 6.3 Analysis methods for RNA-seq data 106
Figure 6.4 Model to deal with hub compound; Note: u,v,x,y denotes pathway; E,F,G,H denotes enzymes.. 110
Figure 6.5 Histogram of length of pathways in our database....................................... 121
Figure 6.6 Histogram for missing enzyme ratio in our pathway database 122
Figure 6.7 Model for VTE2 mutant in *Arabidopsis* 123
Figure 6.8 Vitamin E level for wild type and VTE2 mutant in *Arabidopsis* [194] 126
Figure 6.9 Functional roles of ICS. phylloquinone (B) and SA accumulation following UV induction (C) [200].. 127
Figure 6.10 Accumulation of Camalexin in Leaves of *Arabidopsis* Col-0 Plants, NahG Plants (control), and sid (ICS) Mutant [199]. 127
Figure 6.11 pathway model for ICS (SID2) mutant 128

Figure 7.1 Trichomes on spearmint leaf. a:Non glandular hairy trichome, b:Peltate glandular trichome (PGT), c: Capitate glandular trichome.................................. 135
Figure 7.2 The studied tissue for RNAseq strategy.. 136
Figure 7.3 Quality control for RNA seq result (box plot for each position in read) x-axis: each base in read (bp) y-axis: quality score for each base/position (20: base accuracy is 99%, 30: base accuracy is 99.9%) 138
Figure 7.4 Enrichment GO items by hypergeometric test. X-axis: log(1/p-value) a) Enrichment GO for trichome tissue of spearmint b) enrichment GO for leaf tissue of spearmint... 140
Figure 7.5 Heatmap for different tissue in spearmint and stevia samples................... 141
Figure 7.6 In vitro enzymatic assays of recombinant MsTPSs. GST-tagged MsTPS recombinant enzymes were purified by glutathione-based affinity chromatography and used for in vitro assays with GPP or FPP as substrate. The final products were analysed by GC-MS.. 142
Figure 7.7 GC-MS result for spearmint sample 144
Figure 7.8 Q-PCR verification for several enrichment pathway predicted by our model 149

Chapter 1

INTRODUCTION

Next-generation sequencing platforms are revolutionizing life sciences. Since first introduced to the market in 2005, next-generation sequencing technologies have had a tremendous impact on genomic research. Next-generation technologies have been used for standard sequencing applications, such as genome sequencing and resequencing, and for novel applications, such as molecular marker development by single-nucleotide polymorphisms (SNPs), metagenomics and epigenomics.

Plants are the primary source of calories and essential nutrients for billions of individuals globally [1]. In addition, plants are also a rich source of medical compounds, many of which are still used directly, or as derivatives, to treat a wide range of diseases for humans. Plant-derived compounds are called as metabolites, which can be categorized either as primary metabolites, necessary for maintenance of cellular functions, or as secondary metabolites that are not essential for plant growth and development but are involved in plant biotic and abiotic stress response and plant pollination.

Next-generation sequencing has been widely used for understanding plant metabolisms. By using next-generation sequencing, draft genomes for unknown species and markers for economically-relevant plants for breeding can be generated. New noncoding transcripts (long noncoding RNA) and new mRNAs encoding enzymes can also be obtained and identified easily. For example, the generation of a draft genome for soybean has been used to study oil production with the aim to improve oil yield [2], genome resequencing for soybean and rice has been done to explore genetic diversity [3, 4], and transcriptome data from various plants have been generated to study the production of secondary metabolites [5-7].

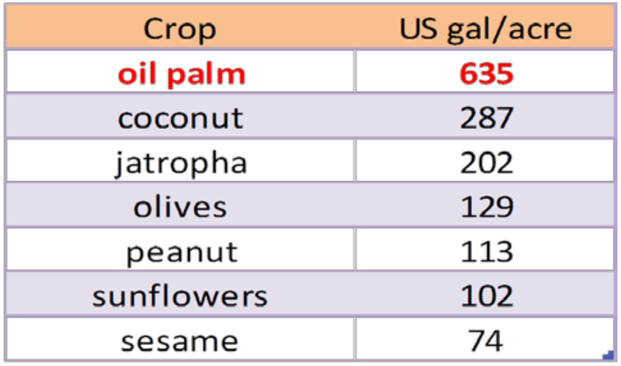
In this thesis, we present several studies where next-generation sequencing has been applied to investigate plant metabolism, with a major focus on lipid and secondary metabolite production. The aim of these studies are: 1) to understand biosynthesis of different plant metabolites, and 2) to increase metabolite production using data generated by next-generation sequencing.

1.1 Motivation

1.1.1 Lipids

Lipids, a major class of primary metabolites, also called fat/oil at room temperature, are an essential component of the human diet. Many plant seeds accumulate storage products during seed development to provide nutrients and energy for seed germination and seedling development. Together, these oilseed crops account for 75% of the world vegetable oil production. These oils are used in the preparation of many kinds of food, both for retail sales and in the restaurant industry. Among these oil crops, oil palm is the most productive in the world’s oil market [Table 1.1]. However, despite being the highest oil-yield crop, whole-genome sequences and molecular resources available for oil palm are very scarce.

Table 1.1 Oil production per weight for oil crops [Wikipedia]



Crop	US gal/acre
oil palm	635
coconut	287
jatropha	202
olives	129
peanut	113
sunflowers	102
sesame	74

Lately large areas of forest are being destroyed to increase the planting areas for oil palm. A better strategy would be to increase the palm fruit/seed oil content. To increase palm fruit/seed oil content, there are two common methods: molecular genetic methods and marker-based breeding.

Although several lipid-related genes/miRNAs have been successfully cloned and investigated in *Arabidopsis* [8], *soybean* [9] and *Jatropha* [10], reports of similar genes in oil palm are still very limited. One major reason is the lack of genome and transcriptome information. Another reason is that it takes a long time to generate transgenic oil palm.

Apart from molecular genetic methods, during the past thirty years, modern breeding methods based on quantitative genetics theory have been extremely successful in improving oil productivity in oil palm. Discovery of the single-gene inheritance for shell thickness and subsequent adoption of D (Dura) X P (Pisifera) planting materials saw a quantum leap in oil-to-bunch ratio from 16% (Dura) to 26% (Tenera). Even with the development of next-generation sequencing, it still remains a big challenge to identify the most common alleles at various polymorphic sites in the oil palm genome and provide data and suggestion for future breeding.

1.1.2 Secondary metabolism

Unlike primary metabolites, secondary metabolites are not involved in essential functions of plants. They typically mediate the interactions of plants with other organisms, such as plant-pollinators, plant-pathogens and plant-herbivores.

Secondary metabolites produced by plants have important uses for humans. They are widely used in pharmaceuticals, flavors, fragrances, cosmetics and agricultural chemical industries [11].

Despite the wide commercial application of secondary metabolites, many of them are produced in low quantities by the plant. Many of these plants have become endangered because of overexploitation.

In the past, genes involved in plant metabolism were often discovered by homology-based cloning [12, 13]. Now, next-generation sequencing technologies have provided an opportunity to scientists to simultaneously investigate thousands of genes in a single experiment. Therefore, new genes/specific transcripts can be discovered and analyzed on a genome-wide basis [14, 15], even without a reference genome. Previous works based on transcriptome analysis have mainly focused on known enzymes and pathways [16, 17], making these methods applicable to some specific plants and known biosynthetic pathways. However, prediction of new functions/metabolic pathways for a plant is still a challenge.

1.1.3 Research challenges

Next-generation sequencing has a lot of applications in modern plant research.

With regard to oil palm research, although recently a draft genome for pisifera oil palm has been released [18], there are still several challenges for the oil palm community:

- The released genome is constructed by a de novo assembly method with 229 different insert libraries. However, it still remains a challenge to assemble other strains of oil palm with a lower coverage, using this released genome.
- It is very important to investigate the genetic variation and diversity during the evolution of oil palm. By identifying polymorphic sites in the genome, key breeding markers can be selected for improving oil yield. Hence, it is necessary to do resequencing work for other commercial oil palm strains to explore their evolutionary history and identify SNP-based markers.
- Identify specific lipid-related genes for oil palm and use the derived sequence information to improve oil yield by molecular genetic approach.
- Build a comprehensive database of the oil palm genome and transcriptome information to be used by biologists.

For secondary metabolism studies, most of the work mainly focuses on known genes/pathways. In the past years, a lot of computational methods on pathway-level analysis have been developed, such as over-representation analysis

(ORA) [19, 20], direct-group analysis [21-23], network-based analysis [24, 25] and model-based analysis [26]. Almost all of these methods try to use enzyme expression levels to select part or all components of specific pathways for a mutation or a treatment. However, these works still share some weaknesses in using enzyme expression level:

- All pathways are considered independent by these methods, which may be not reasonable. They apply the raw expression level of enzymes for each pathway, although some enzymes/compounds may be involved in more than one pathway.
- Many major secondary metabolite-related plants do not have a reference genome. Consequently, many enzymes in reference pathways are missing. This missing information makes applying these methods challenging.

1.2 Thesis contribution

Next-generation sequencing is a useful tool for studying plant metabolisms. In our study, we focus on lipid and secondary metabolism. For the lipid study, we first develop a novel reference-based genome assembly pipeline and apply it to assemble the genome of dura oil palm. Then, we investigate the evolutionary history and genetic variation of oil palm by reseqeuncing 12 different oil palm strains. Lastly, an online database is built to visualize genome information for oil palm. For the secondary metabolism study, we introduce a novel weighted pathway approach and use it to predict new functions/metabolic pathways for the plants studied.

Specifically:

- We generate different genomic libraries for dura oil palm using next-generation sequencing techniques.
- We propose a comprehensive reference-based genome assembly pipeline, which performs mis-assembled scaffold identification and repeat scaffold identification.
- We resequence 12 different oil palm strains from all over the world.
- We explore the evolutionary history and genetic variation between different oil palm strains.
- We build a database and a blast tool to show and visualize genome information for oil palm.
- We propose a weighted pathway approach, which takes into account the dependency between different pathways.
- We validate our weighted pathway approach on mint samples (leaf, leaf without trichome and trichome tissue), and predict some new functions/metabolic pathways for mint.

1.3 Thesis organization

The rest of this thesis is organized as follows. Chapter 2 presents some background and related work for next-generation sequencing study. Chapter 3 gives details of our reference-based genome assembly pipeline. Chapter 4 presents how to apply this reference-based genome assembly pipeline to construct a draft genome for Dura oil palm. Chapter 5 describes the database and blast tool for oil palm genome resource. Chapter 6 discusses the weighted pathway approach. Chapter 7 describes how to apply the weighted pathway approach on mint samples. Chapter 8 gives a summary of the work and proposes some future research directions.

1.4 Declaration

This dissertation is based on the following material:

- Jingjing Jin, May Lee, Jian Ye, Rahmadsyah, Yuzer Alfiko, Chin Huat Lim, Antonius Suwanto, Zhongwei Zou, Bing Bai, Limsoon Wong, Gen Hua Yue , and Nam-Hai Chua: The genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm, in preparation. (Chapter 3 and Chapter 4)
- Jingjing Jin, Jun Liu, Huan Wang, Limsoon Wong, Nam-Hai Chua: PLncDB: plant long non-coding RNA database. Bioinformatics 2013, 29:1068-1071. (Chapter 5)
- Jingjing Jin, Qian Wang, Haojun Zhang, Hufeng Zhou, Rajani Sarojam, Nam-Hai Chua and Limsoon Wong: Investigating plant secondary metabolisms by weighted pathway analysis of next-generation sequencing data, in preparation. (Chapter 6)
- Jingjing Jin, Deepa Panicker, Qian Wang, Mi Jung Kim, Jun Liu, Jun -Lin Yin, Limsoon Wong, In-Cheol Jang, Nam-Hai Chua and Rajani Sarojam: Next generation sequencing unravels the biosynthetic ability of Spearmint (Mentha spicata) peltate glandular trichomes through comparative transcriptomics, BMC Plant Biology, 2014, accepted. (Chapter 7)

- Jingjing Jin, Mi Jung Kim, Savitha Dhandapani, Jessica Gambino Tjhang, JunLin Yin, Limsoon Wong, Rajani Sarojam, Nam-Hai Chua and In-Cheol Jang: Floral transcriptome of Ylang Ylang (Cananga odorata var. fruticosa) uncovers the biosynthetic pathways for volatile organic compounds and a multifunctional and novel sesquiterpene synthase, Journal of Experimental Botany, submitted. (Chapter 7)

Chapter 2

RELATED WORK

2.1 Next-generation sequencing

Next-generation sequencing (NGS) techniques became commercially available around 2005, the first one being the Solexa sequencing technology [27]. Since then, several different methods have been developed, which can largely be grouped into three main types: sequencing by synthesis, sequencing by ligation and single-molecule sequencing.

Sequencing by synthesis involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with a chemiluminescent enzyme [28]. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base at a time, and detecting which base is actually added at each step. The well-known methods in this group include 454, Illumina and Ion Torrent, differing by read length and template method [Table 2.1].

Table 2.1 Comparison of performance and advantages of various NGS platform [27]

Platform	Library	length	#Read	output	accuracy	Run time	cost (US$)	Pros	Cons
Sequencing by synthesis									
Roche/454	Frag, MP/emPCR	700	~1 million	700 Mb	100.00%	23h	500,000	Long reads, fast run times;	Higher reagent costs, low error rates
Illumina HiSEq 2000	Frag, MP, solid-phase	2 × 100	>5 million	~570 Gb	>80 % > Q30	8.5d	600,000	Currently most widely used platform, high coverage	Shorter read lengths
Ion Torrent PGM	Frag, emPCR	200	5 million	1 Gb	99.99%	2h	50,000	Very fast run time, cost effective	low throughput
Sequencing by ligation									
Life/AB SOLiD 5500 Series	Frag, MP/emPCR	75 × 35	~1 billion	~120 Gb	99.99%	7d	600,000	2-Base encoding error correction	Longest run times
Polonator G.007	MP only/emPCR	26	~80 million	5–12 Gb	>98 %	5d	170,000	Open source; cost effective	Users maintain; shortest NGS lengths
Single-molecule sequencing									
Helicos BioSciences HeliScope	Frag, MP/ single-molecule	35	~1 billion	35 Gb	99.995	8d	999,000	High multiplexing ability,no template amplification	Short read lengths, high error rates
Pacific BioScience PacBio HRS	Frag only/ single-molecule	1300	35000	45 Mb	100.00%	1h	700,000	Longest reads, no template amplification	Highest error rates

Sequencing by ligation is a type of DNA sequencing method that uses the enzyme DNA ligase to identify the nucleotide present at a given position in a DNA sequence. Unlike sequencing-by-synthesis methods, this method does not use a DNA polymerase to create a second strand. Instead, the mismatch sensitivity of a DNA ligase enzyme is used to determine the underlying sequence of the target DNA molecule [27]. SOLiD and Polonator belong to this group; they differ in their probe usage and read length.

Single-molecule sequencing (SMS), often termed “third-generation sequencing”, is based on the sequencing-by-synthesis approach. The DNA is synthesized in zero-mode wave-guides (ZMWs), which are small well-like containers with the capturing tools located at the bottom of the well. The sequencing is performed with the use of unmodified polymerase (attached to the ZMW bottom) and fluorescently labeled nucleotides flowing freely in the solution. This approach allows reads of 20,000 nucleotides or more, with an average read length of 5k bases, such as Pacific BioScience's technique [Table 2.1]. SMS technologies are relatively new to the market, and in future will become more readily available.

NGS technologies are evolving at a very rapid pace, with established companies constantly seeking to improve performance, accessibility and accuracy, such as nanopore sequencing [29], which is based on the readout of electrical signals occurring at nucleotides passing by alpha-hemolysin pores covalently bound with cyclodextrin.

The various NGS platforms currently available or under development have different methods to sequence DNA, each employing various strategies of template preparation, immobilization, synthesis and detection of nucleic type and order [27]. These methodological differences produce different sequencing result, such as read length, throughput, output and error rates, with each platform having important advantages and disadvantages [Table 2.1]. Nevertheless, next-generation sequencing technologies are paving the way to a new era of scientific discovery. As sequencing techniques become easier, more accessible, and more cost effective, genome sequencing will become an integral part of every branch of the life sciences; plant biology is no exception. Hence, in sections below, we summarize the special usage of next-generation sequencing in plant biology.

2.2 Whole-genome sequencing

It is not surprising that considerable effort has been given to the sequencing of plant genomes during the last decades. The dissected genomes enable the identification of genes, regulatory elements, and the analysis of genome structure [30]. This information facilitates our understanding of the roles of genes in plant development and evolution, and accelerates the discovery of novel and functional genes related to biosynthesis of plant metabolites. Reference genomes are also important in the identification, analysis and exploitation of the genetic diversity of an organism in plant population genetics and breeding studies [30].

The first completed reference genomes in plants, *Arabidopsis* [31], was a major milestone not only for plant research but also for genome sequencing. The approach relied on overlapping bacterial artificial chromosomes (BAC) clones that represent a minimal tiling path to cover each chromosome arm. The BAC sequences were individually assembled and arranged according to the physical map, creating a genome sequence of very high quality. The high effort and time associated with this approach limited its applicability only to a few plant genomes. Nevertheless, after three years, the first crop plant, rice, was also constructed based on the BAC approach [32, 33].

Next, many groups adopted an alternative strategy: whole-genome sequencing (WGS). In WGS method, a whole genome is randomly broken down into small pieces, which are then sequenced and subsequently assembled. This method has been improved with the use of multiple libraries of different insert sizes. The first WGS efforts were mainly implemented on smaller genomes, including Poplar [34], Grape [35] and Papaya [36]. These sequencing methods are called first-generation sequencing techniques (mainly using Sanger-based methods). Further refinement on the WGS approach enables the sequencing of larger genomes, such as Sorghum bicolor [37] and soybean [2]. Compared to BAC-based methods, time and cost of these projects are reduced a lot. However, the reduction in time and cost is achieved at the expense of assembly fidelity in repetitive regions and expanding need for computer hardware resources. Although WGS reduced the time and effort requirement, genome sequence generation was still expensive and time consuming, due to the high cost of Sanger sequencing.

The use of next-generation sequencing (NGS) platforms in WGS projects improved the output and cost ratio of sequencing dramatically. The application of NGS to plant genomes has become an increasingly strong trend. Although several plant genomes were generated by combination of NGS with Sanger sequencing

[38, 39], more and more genomes were sequenced using NGS alone. More recently, Illumina sequencing emerged as the dominant NGS platform for genome sequencing, providing data pools for recent genomes such as Chinese cabbage [40], potato [41], orange [42], banana [42] and watermelon [43].

Despite the advancement of genome sequencing technologies, the downstream analysis of short-read datasets after sequencing is a tough task; one of the biggest challenges for the analysis of high-throughput sequencing reads is whole-genome assembly. As genome sequencing technologies evolve, methods for assembling genomes have to keep step with them.

At the beginning, although the output was limited, the length of sequencing reads was much longer (~460bp for the first published genome). Several assemblers have been developed to assemble genomes from these long (“Sanger”) reads, including the Celera Assembler [44], ARACHNE [45] and PCAP [46]. These algorithms assemble the reads in two or more distinct phases, with separate processing of repetitive sequences. First, they assemble reads with unambiguous overlaps, creating contigs that end on the boundaries of repeats. Then, in a second phase, they assemble the unambiguous contigs together into larger sequences, using mate-pair constraints to resolve repeats. They are called Overlap/Layout/Consensus (OLC)-based assembly methods, which try to connect each read by overlap. More recently, the Newbler [47] assembler has been specifically designed to handle 454 Life Sciences (Roche) reads, which have a different error profile from that of Sanger long reads.

In principle, assemblers created for long reads can also facilitate assembly of short reads. The principles of detecting overlap and building contigs are no different. In practice, initial attempts to use previous assemblers for very short reads, which are mostly generated by next-generation sequencing platforms, either failed or performed very poorly, for a variety of reasons. Some of these failures were easy to understand: for example, assemblers impose a minimum read length, or they require a minimum amount of overlap, which may be too long for a short-read sequencing project. Another problem is that the computation of overlaps is one of the most critical steps in long-read assembly algorithms. Short-read sequencing projects may require a redesign of this step to make it computationally feasible, especially since many more short reads are generated by next-generation sequencing platforms than long-read platforms. For these reasons and others, a new group of genome assemblers has been developed specifically to address the challenges of assembling very short reads. These assemblers include Velvet [48], ALLPATHS [49], ABySS [50], Gossamer [51], oases [52], SparseAssembler [53], IDBA [54] and SOAPdenovo [55]. Different from using an overlap graph, all of these assemblers are based on de Bruijn graph. In these approaches, the reads are decomposed into k-mers that in turn become the nodes of a de Bruijn graph. A directed edge between nodes indicates that the k-mers on those nodes occur consecutively in one or more reads. These k-mers take the place of the seeds used for overlap computation in assemblers for long reads. However, at times, the cost of genome sequencing or the biological properties of a genome sequence compels a genome to be sequenced at a lower coverage. Since most plant genomes are large, cost is still a major factor. Hence, relatively few plant species have been sequenced, compared with the hundreds of thousands of species around the world, especially for plants with large genome.

Recently, as more and more reference genomes have been released, there is a widespread interest in sequencing large numbers of closely related species or strains, by relatively low coverage sequencing. This can help in exploring population structure and genetic variation. By aligning the de novo assembly scaffolds to a reference genome---thus ordering and orientating the scaffolds---the assembly results can be considerably improved. This process/method is called reference-based genome assembly; examples include ABACAS [56], PAGIT [57], RACA [58] and eRGA [59]. It is a useful technique for genome assembly, due to a lower sequencing depth requirement of the target genome.

Sequencing is a rapidly advancing field, and third-generation sequencing technologies have already announced some features with even longer read and insert sizes. The use of new sequencing methods and technologies will expand our knowledge of plant genomes and contribute to plant genetics.

2.3 Genome resequencing

With the development of next-generation sequencing technologies, reference genome sequences for many plants are available, cataloguing sequence variations and understanding their biological consequences have become a major research aim. However, for large eukaryotic genomes such as human or different plants, even high-throughput sequencing technologies can only allow deep genome-wide sequence coverage of a small number of individuals. However, resequencing the genome of many individuals for which there is a reference genome allows investigation of the relationship between sequence variation and normal or disease phenotypes. When the new sequencing power is targeted to limited areas of large genomes [60], it is feasible to study variation in specific regions in thousands of individuals.

By resequencing 50 strains of cultivated and wild rice, molecular genetic analyses indicated that *indica* and *japonica* originated independently. Meanwhile, population genetics analyses of genome-wide data of cultivated and wild rice have also suggested that *indica* and *japonica* genomes generally appear to be of independent origin [3].

Another successful application in plants is the resequencing of 31 wild and cultivated soybean genomes [4], which has identified a set of 205,614 tag SNPs for QTL mapping and marker development.

For domestic animals, such as chicken [61], by whole-genome resequencing, many potential selection loci were found to play important roles during evolution, which provided some good evidence for future breeding of domestic animals.

Increasingly, powerful sequencing technologies are reaching an era of individual/personal genome sequences and raising the possibility of using such information to guide breeding or medical decisions. Genome resequencing also promises to accelerate the identification of disease-associated mutations in plants or human. More than 80% of a typical mammalian genome is composed of repeats and intergenic or noncoding sequences [5]. Thus, in the future, it is crucial to focus resequencing only on high-value genomic regions. Protein-coding exons represent one such type of high-value target by many groups, which are commonly called exome sequencing [62].

2.4 Molecular marker development

Linkage mapping and evolutionary studies in plants rely on the power of identifying and understanding single-nucleotide and insertion-deletion polymorphisms (SNP), which can reflect the differences in a phenotype of interest. This is an important approach in improving the yield of crop plants.

Previous implementation of high-throughput PCR-based marker technologies and introduction of first-generation sequencing, such as Sanger sequencing, have increased the number of markers as well as the individuals in marker-based studies [27]. These new changes enabled a new era in linkage mapping analysis and breeding studies in plants, which is called marker-assisted selection (MAS).

More recently, next-generation sequencing technologies have enabled genome-wide discovery of SNPs on a massive scale. The 454 platform has some successful applications on maize for SNP discovery [63]. However, the higher throughput and lower cost of Illumina and SOLiD technologies have made them much more popular for major programs when a reference genome is available [64]. Even for plant species where high-quality reference genomes are not available [65, 66], some reference-free based variant calling methods have been developed to deal with them, such as high-quality transcriptome assembly results or some de novo partial assemblies from BAC contigs (chapter 2.2).

Another important family benefitting from NGS is simple sequence repeats (SSRs or microsatellites), which are repeating DNA sequences (tandem arrays) of 1-6 nucleotides that occur in all prokaryotic and eukaryotic genomes. Their high mutation rate and polymorphism, multi-allelic and co-dominant nature, and need for little DNA for gathering data, make them a good choice for various applications, such as linkage map development, quantitative trait loci (QTL) mapping, marker-assisted selection, genetic diversity study and evolution study [27, 28]. Previously, SSRs were developed by constructing genomic libraries using recombinant DNA enriched for a few targeted SSR motifs, followed by isolation and sequencing of clones containing SSRs [27]. Based on NGS, sequence of more and more genomes for plant species have been determined, which enables the discovery of potential SSRs just by de novo searching on the genomes. Zalapa et al. showed the power of NGS for developing SSRs in plants through a review of their work in strawberry and 95 other studies by next-generation sequencing platforms [67].

2.5 Transcriptome sequencing

The sequencing of DNA products (cDNA), which are synthesized from mRNA isolates, have played important roles in gene expression analysis, discovery and determination of alternative splicing forms of genes (isoforms). For a species with a genome available, cDNA sequencing can facilitate the annotation of splicing sites, transcribed regions in the genome (such as long noncoding RNA), as well as improve gene prediction algorithms [68].

More recently, the increasing gains from next-generation sequencing techniques, as well as improvement in short-gun RNA sequencing (RNA-seq) strategies, have provided relatively high coverage for gene discovery, annotation and polymorphism discovery in both model and non-model plant species, which are rapidly replacing other methods of studying gene expression such as microarrays. It is practical in non-model plants, because reference genomes are not required by RNA-seq. Similar to algorithms used for genome assembly, several tools, including Trinity [69] and Oases [52], have been developed for RNA-seq assembly, although they have slight differences in dealing with alternative splicing. Afterwards, many new genes and transcription factors (TFs) have been identified to play roles in plant metabolite biosynthesis [6, 70].

Different from gene-level analysis, some people attempt to shift from analysis of individual genes to a set of genes, which perform a specific function together [71]. In the past decade, the knowledge which describes---using the standardized nomenclature of GO terms---the biological processes, components, and molecular functions in which individual genes and proteins are known to be involved in, as well as---using the not-so-standardized nomenclature of biological pathways---how and where gene products interact with each other, have expanded dramatically. Therefore, based on transcriptome expression level by RNA-seq, some researchers attempt to analyze them at the functional level. They try to identify interesting GO terms or pathways of specific tissue or treatment. These methods include: over-representation analysis (ORA) [72] which identifies enriched GO terms/pathways based on a list of differentially expressed genes, direct-group analysis [73, 74] which assigns different scores for different GO terms/pathways, network-based analysis [24, 25] which identifies in each pathway a subset of genes most relevant to a phenotype, and model-based analysis [26, 75] which uses dynamic models of pathways to identify aberrant pathways in a phenotype. Although each of these different methods has its own advantages/disadvantages and scope, most of them have some successful applications in plant metabolism research.

2.6 Non-coding RNA characterization

RNAs in eukaryotic cells can be classified into five categories: ribosomal RNAs (rRNA), transfer RNAs (tRNAs), messenger RNAs (mRNAs), long noncoding RNAs (lncRNAs) and small RNAs (sRNAs). Over 90% of the total RNA molecules present in a cell are rRNAs and tRNAs, while sRNAs account for ~1% or less. Eukaryotic regulatory sRNAs are a subset of sRNAs ranging in size from ~20 to 30nt; they include microRNAs (miRNAs), small interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs). The functions of these regulatory sRNAs are conserved from plants to animals, which imply their involvement in fundamental cellular processes. Discovery and profiling of these regulatory sRNAs are of primary interest in unraveling their regulatory functions.

In the past, various experimental methods---including cloning, Northern blot, RNase protection assay and primer extension---have been applied to quantify and identify novel small RNAs. After the discovery of the fold-back structure characteristic of lin-4 and let-7 [76], many small RNAs were identified by cloning and sequencing. Although cloning and sequencing is a very useful method for the identification of individual novel miRNAs, there are still limitations for this method. First, it requires a lot of total RNA, which is not practical in many cases. In addition, due to low coverage, some small RNAs with low abundance may be missed. Sometime, it is very difficult to distinguish between miRNAs and other ncRNAs, rRNAs or tRNAs. To avoid these limitations, many researchers have adopted Northern blotting analysis [77], which can efficiently detect miRNAs. RNase protection assays are mainly used to detect mature miRNAs [78].

Microarray technology is a further step toward high-throughput quantification of miRNA expression, and it has also been used to explore miRNA expression in various tissues and development stages [79]. A good case is miRNA microchip, which is specifically designed for miRNA profiling on a global level [80]. Compared with other experimental methods, miRNA-specific arrays have several advantages. First, the expression of multiple RNAs can be detected and measured at the same time. Second, the expression of mature and precursor miRNAs can be detected simultaneously by some careful probe design strategy. In addition, less amount of RNA is needed, when compared to that required for other experimental methods, such as Northern blot.

Although cloning and sequencing of small RNAs can discover novel miRNAs, it is time consuming and limited to the most abundant small RNAs. Real-time PCR enables rapid detection of miRNAs and their precursors, but has limitations on novel miRNA identification. miRNA-related arrays also have limitations on novel miRNA identification. In contrast, high-throughput sequencing not only revolutionizes mRNA discovery, but also accelerates the discovery of small RNAs and reveals their expression patterns. For species with a known reference genome, just by mapping and structure checking, many known and novel small RNAs can be easily detected. For example, using the Solexa platform, the NK cell miRNA transcriptome has been investigated to study miRNA roles in NK cell biology, and 21 novel miRNA genes have been discovered [81]. Using the Illumina platform, novel miRNAs, phased smRNA clusters and small-interfering RNAs have been identified in *Arabidopsis* [82].

Therefore, with the development of small RNA sequencing, many associated bioinformatics software and tools---e.g., miRDeep [83], UEA small RNA tools [84]---have been developed to identify known and novel miRNAs with sequencing reads and reference genomes. Particularly, for plants whose genome information is unavailable, small RNA sequencing shows remarkable superiority over other methods. This is because the small RNA reads can be mapped to public small RNA database to identify the known small RNAs. However, it is still a challenge to identify novel miRNAs for these species.

Apart from small-RNA profiling, identification of long noncoding RNAs also benefits greatly from next-generation sequencing. Some researchers attempt to detect long noncoding RNAs by identifying trimethylation of lysine 4 of histone H3 (H3K4me3) peaks at their gene promoter and trimethylation of lysine 36 of histone H3 (H3K36me3) peaks along the length of the transcribed gene region based on CHIP-seq technique [85]. However, most researchers employ RNA-seq to detect long noncoding RNAs using the hypothesis that all un-annotated transcripts in the genome, which can be transcribed, but not translated, could be considered as potential long noncoding RNAs. Using RNA-seq, the transcribed regions in the genome can be found easily, which are good candidates for long noncoding RNAs.

As NGS technologies continue to improve, their scope and application will correspondingly expand within and across scientific research. Plant biology has gained much from increasing capacity in genomics, plant breeding, evolutionary studies and biosynthesis of different products/metabolites. In this thesis, we introduce several studies to understand plant metabolism using next-generation sequencing techniques in following chapters.

Chapter 3

REFERENCE-BASED GENOME ASSEMBLY

In Chapter 2, we have mentioned that considerable effort has been devoted to the sequencing of plant genomes during the last two decades. This is because a sequenced genome enables the identification of genes, regulatory elements, and the analysis of genome structure [30]. Moreover, this information facilitates our understanding of the roles of genes in plant development and evolution, and accelerates the discovery of novel and functional genes related to biosynthesis of plant metabolites.

The development and commercialization of next-generation massively parallel DNA sequencing technologies—including Illumina's Genome Analyzer (GA) [86], Applied Biosystems' SOLiD System, and Helicos BioSciences' HeliScope [87]—have revolutionized genomic research. The use of next-generation sequencing (NGS) platforms in whole-genome sequencing projects has improved the output and cost ratio of sequencing dramatically. The application of NGS to plant genomes has become an increasingly strong trend.

In the past two decades, as genome sequencing technologies evolve, methods for assembling genomes have also considerably evolved alongside.

3.1 Background

According to the scope and theory, NGS assemblers are commonly classified into two major categories: Overlap/Layout/Consensus (OLC)-based assembly methods and de Bruijn Graph (DBG)-based assembly methods.

3.1.1 OLC-based assembly methods

In the traditional approach, assembly is formalized using the overlap graph. This structure represents each sequencing read as a separate node, where two reads presenting a clean overlap are connected by a directed edge. These algorithms assemble the reads in two or more distinct phases, with separate processing of repetitive sequences. First, they assemble reads with unambiguous overlaps, creating contigs that end on the boundaries of repeats. In the second phase, they assemble the unambiguous contigs into longer sequences, using mate-pair constraints to resolve repeats. Newbler (454/Roche), ARACHNE [45], Edena [88] and SGA [89] belong to this category of methods. They are called Overlap/Layout/Consensus (OLC)-based assembly methods, which try to connect each read by overlap.

However, this approach has two serious shortcomings that make it applicable for long-read sequencing only, like those produced by 454 sequencing technique. Firstly, the link of the two reads is determined by the overlap nucleotide sequence, and this overlap has to be sufficiently long to ensure a reliable link. For example, in a study by Narzisi and Mishra [90], they found that compared to other de novo assembly methods, an OLC-based method---Edena---not only produced smaller N50 size, but also a larger number of total scaffolds on a short-read dataset for a known genome [Table 3.1]. Hence, this method is only applicable to long reads, not applicable to short sequences, such as those produced by Illumina sequencing.

Table 3.1 Comparison between different assemblers on short reads example for a known genome [90]

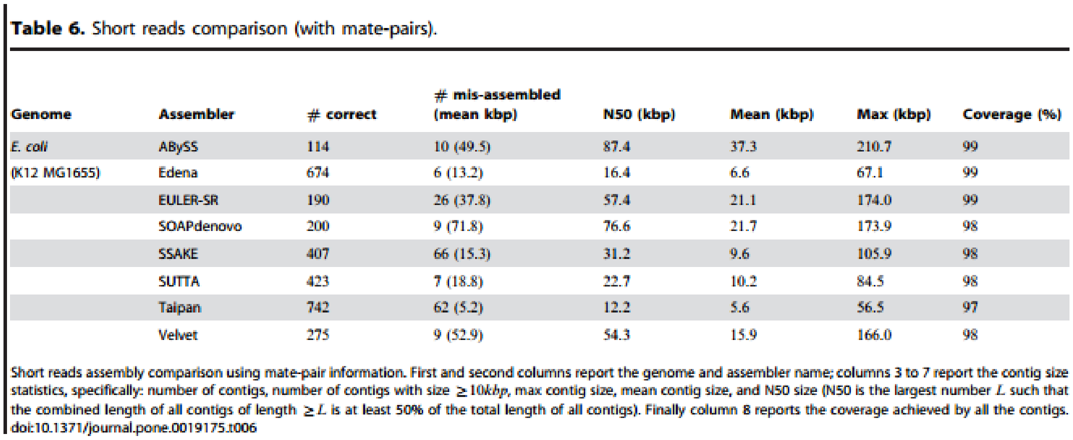


Table 6. Short reads comparison (with mate-pairs).

Genome	Assembler	# correct	# mis-assembled (mean kbp)	N50 (kbp)	Mean (kbp)	Max (kbp)	Coverage (%)
E. coli	ABySS	114	10 (49.5)	87.4	37.3	210.7	99
(K12 MG1655)	Edena	674	6 (13.2)	16.4	6.6	67.1	99
	EULER-SR	190	26 (37.8)	57.4	21.1	174.0	99
	SOAPdenovo	200	9 (71.8)	76.6	21.7	173.9	98
	SSAKE	407	66 (15.3)	31.2	9.6	105.9	98
	SUTTA	423	7 (18.8)	22.7	10.2	84.5	98
	Taipan	742	62 (5.2)	12.2	5.6	56.5	97
	Velvet	275	9 (52.9)	54.3	15.9	166.0	98

Short reads assembly comparison using mate-pair information. First and second columns report the genome and assembler name; columns 3 to 7 report the contig size statistics, specifically: number of contigs, number of contigs with size $\geq 10kbp$, max contig size, mean contig size, and N50 size (N50 is the largest number L such that the combined length of all contigs of length $\geq L$ is at least 50% of the total length of all contigs). Finally column 8 reports the coverage achieved by all the contigs.
doi:10.1371/journal.pone.0019175.t006

Secondly, the computation of pairwise overlaps is inherently quadratic in complexity, although it can be optimized by heuristics [91] and filters [92]. For short-read sequencing, several hundred million reads are typically produced. Thus this quadratic time complexity is not acceptable.

In summary, due to the large-size requirement for the reads and computation time limitation, methods based on this approach are only applicable for low-throughput long-read sequencing datasets.

3.1.2 DBG-based assembly methods

In 1995, Idury and Waterman [93] introduced the use of a sequence graph to represent an assembly. They presented an assembly algorithm for an alternative sequencing technique, sequencing by hybridization, where an oligoarray could detect all the k-nucleotide words, also known as k-mers, present in a given genome. By connecting the nodes (k-mers) corresponding to every detected word, they could produce contigs, which are chains of overlapping k-mers.

Pevzner et al. [94] expanded on this idea. Firstly, they proposed a slightly different formalization of the sequence graph, called a de Bruijn graph, whereby the k-mers are represented as arcs and overlapping k-mers join at their tips. For the k-mers, users can adjust by themselves, which removes the size limitation of overlap-based methods. A lot of software based on de Bruijn graph have now been developed, such as SOAPdenovo [55], SparseAssembler [53], ABySS [50], Velvet [48], oases [52], IDBA [54], Minia [95] and Allpaths LG [49], which use different techniques to deal with repeats and tips. Most of them have been successfully applied to construct the draft genome of different species [42, 43, 96, 97], with more than 100x coverage. These drafts are of high quality, and, although imperfect, have served as references for the community.

However, at times, the cost of genome sequencing or the biological properties of a genome sequence forces a genome to be sequenced at a lower coverage. Since mammalian genomes are large, cost is a major factor. Hence, it is still a challenge for genome assembly with lower sequence coverage.

3.1.3 Reference-based genome assembly

The recent development of ultra-high-throughput sequencing technologies and the sequence assembly approaches mentioned in previous sections have led to a huge increase in the number of genome sequencing projects being carried out [98]. In particular, there is widespread interest in sequencing a large number of closely related species or strains, where a high-quality reference genome already exists, by low-coverage sequencing. This can help in exploring population structure and genetic variation.

By aligning the de novo assembly scaffold to a reference genome---thus ordering and orientating the scaffold---the assembly results can be improved a lot. This process/method is called reference-based genome assembly. It is a useful technique to genome assembly by lowering the sequencing depth requirement of the target genome.

In the past several years, four such assemblers---viz, ABACAS[56], PAGIT[57], RACA [58] and eRGA [59]--- have been developed for carrying out both de novo assembly and mapping assembled scaffolds to reference genome.

eRGA focuses on merging results from de novo assembly and raw-read alignment methods; strictly speaking, it is not a reference-based assembly tool.

ABACAS first conducts a de novo assembly and then aligns the resulting contigs/scaffolds to a reference genome to obtain much longer ones. However, it suffers from several disadvantages:

- it filters scaffolds by the ***Identity*** criteria [***Identity*** criteria: percentage of match region between scaffold and target reference genome], which may discard some mis-assembled scaffolds;
- it randomly assigns repeat scaffolds to the reference genome;
- it is only applicable to a reference genome that has one chromosome.

PAGIT is a unified software based on a series of previous software and, in particular, ABACAS is its main component. RACA tries to construct synteny block between different scaffolds. However, they also don’t deal with mis-assembled scaffolds and overlap scaffolds.

The reference-based genome assembly approach has been successfully used to assemble four different *Arabidopsis* genomes [99]. However, the pipeline used in that study is based on alignment of raw reads to a reference genome, which may miss some rearrangement parts, despite having some extension mechanism to mitigate this problem.

3.2 Methods

In section 3.1.3, we have mentioned that there are several disadvantages in existing reference-based genome assembly methods. To overcome these disadvantages, we propose some new components to form a new pipeline for reference-based genome assembly:

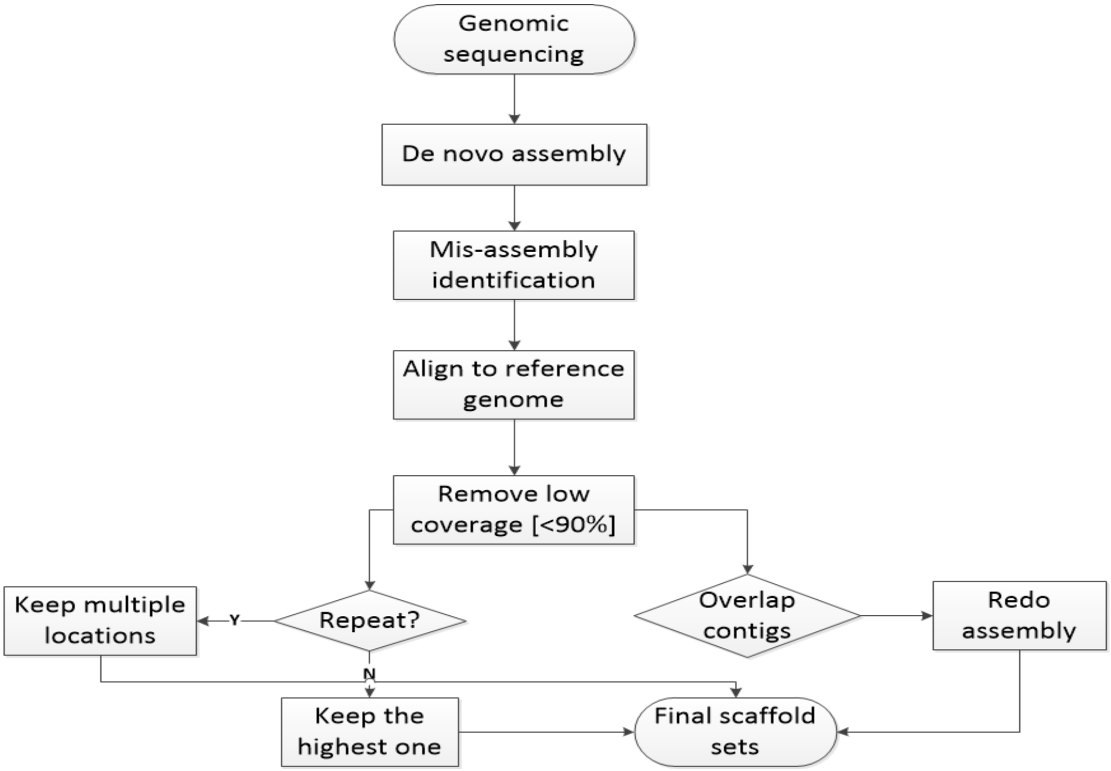
- In order to solve the problem caused by mis-assembled scaffolds, we first try to identify the mis-assembled scaffolds. Then, we correct these mis-assembled scaffolds.

- For the repeat scaffolds, we do not randomly assign them a location in the reference genome. According to a hypothesis to be stated later, some of them are assigned to multiple locations.

- If all chromosomes of a reference genome are merged into one combined sequence, it will cost a lot of memory when doing alignment between scaffolds and the reference genome. Hence, in our proposed pipeline, they will be considered separately.

In summary, we are proposing a much more comprehensive pipeline for reference-based genome assembly [Figure 3.1]. In particular, our proposed pipeline has five components: 1) de novo assembly, 2) mis-assembled scaffold identification and correction, 3) alignment to a reference genome, 4) repeat scaffold identification, and 5) resolution of overlapping scaffolds. We discuss each component of our pipeline in the following sections.

Figure 3.1 Pipeline of our proposed reference-based genome assembly pipeline

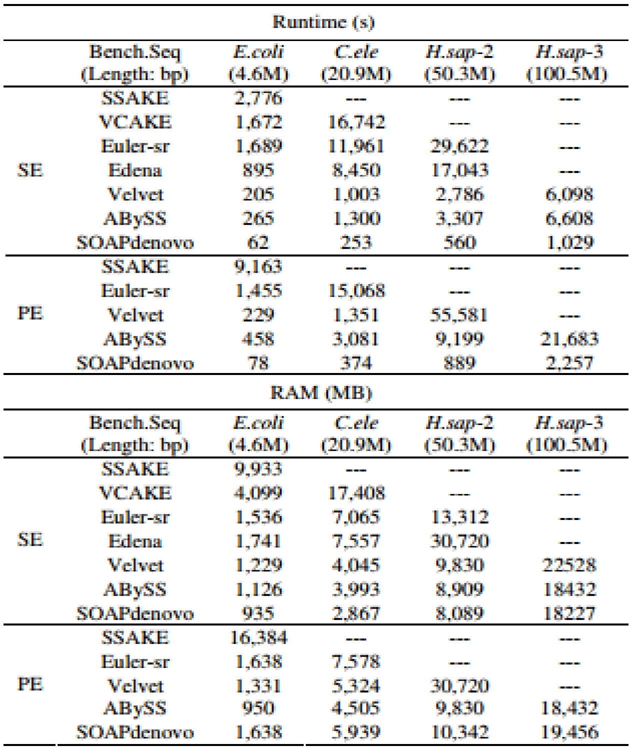


3.2.1 De novo assembly

As introduced in earlier sections, many different de novo assembly tools have been developed for different sequencing results. From a comparison of running time and RAM for different assembly methods [Table 3.2], the DBG-based methods SOAPdenovo [55] and ABySS [50] consume less RAM and time, especially for a large genome [part of human genome with size of 100.5M bp]. Considering this superior performance and the features of our sequencing data sets, these DBG-based methods should be adopted for our analysis.

Hence, both of them and other well-known tools---IDBA [54], Velvet [48], Oases [52], SparseAssembler [53], Gossamer [51] and Allpaths-LG [49], which also have less RAM and time costs---are used to do the first de novo assembly step.

Table 3.2 Comparison of running time (Runtime) and RAM for different de novo assembly method [100]. SE denotes single-end sequencing dataset. PE denotes pair-end sequencing dataset. E.coli, C.ele, H.sap-2, H.sap-3 denotes four different test dataset. Second column denotes different de novo assembly method. ---denotes RAM of the server is not enough or running time too long (>10 days). s denotes second. MB denotes megabytes.



Runtime (s)

	Bench.Seq (Length: bp)	*E.coli* (4.6M)	*C.ele* (20.9M)	*H.sap*-2 (50.3M)	*H.sap*-3 (100.5M)
SE	SSAKE	2,776	---	---	---
	VCAKE	1,672	16,742	---	---
	Euler-sr	1,689	11,961	29,622	---
	Edena	895	8,450	17,043	---
	Velvet	205	1,003	2,786	6,098
	ABySS	265	1,300	3,307	6,608
	SOAPdenovo	62	253	560	1,029
PE	SSAKE	9,163	---	---	---
	Euler-sr	1,455	15,068	---	---
	Velvet	229	1,351	55,581	---
	ABySS	458	3,081	9,199	21,683
	SOAPdenovo	78	374	889	2,257

RAM (MB)

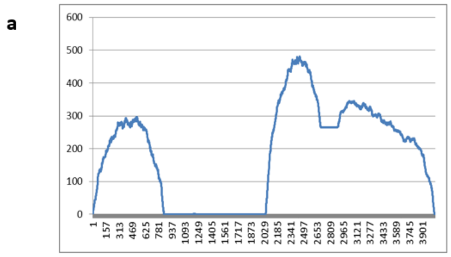
	Bench.Seq (Length: bp)	*E.coli* (4.6M)	*C.ele* (20.9M)	*H.sap*-2 (50.3M)	*H.sap*-3 (100.5M)
SE	SSAKE	9,933	---	---	---
	VCAKE	4,099	17,408	---	---
	Euler-sr	1,536	7,065	13,312	---
	Edena	1,741	7,557	30,720	---
	Velvet	1,229	4,045	9,830	22528
	ABySS	1,126	3,993	8,909	18432
	SOAPdenovo	935	2,867	8,089	18227
PE	SSAKE	16,384	---	---	---
	Euler-sr	1,638	7,578	---	---
	Velvet	1,331	5,324	30,720	---
	ABySS	950	4,505	9,830	18,432
	SOAPdenovo	1,638	5,939	10,342	19,456

Among these approaches, the one producing the longest N50 and larger genome coverage [>85%] is then selected for use in the downstream analysis in our pipeline.

3.2.2 Mis-assembled scaffold identification and correction

Although there are many successful de novo assembly tools [48-50, 54, 55], there are many mis-assembled scaffolds in their output, especially for large genomes, due to sequencing errors and repeat regions. If these mis-assembled scaffolds are not corrected, most of them will be excluded by the ***Identity*** (percentage of length of match region and total length of scaffold) criterion in typical reference-based assembly methods and, thus, negatively impacting the completeness of the resulting assembled genome. Hence, it is necessary to identify and correct mis-assemblies before aligning them to the reference genome in reference-based genome assembly methods.

Figure 3.2 An example of a mis-assembled scaffold [scaffold148]. a. the coverage across the scaffold 148 by insert size of pair end reads b. the detail alignment information for scaffold 148 after aligning to the reference genome. In this figure, t denotes target reference genome, q denotes query assembly scaffolds.



b

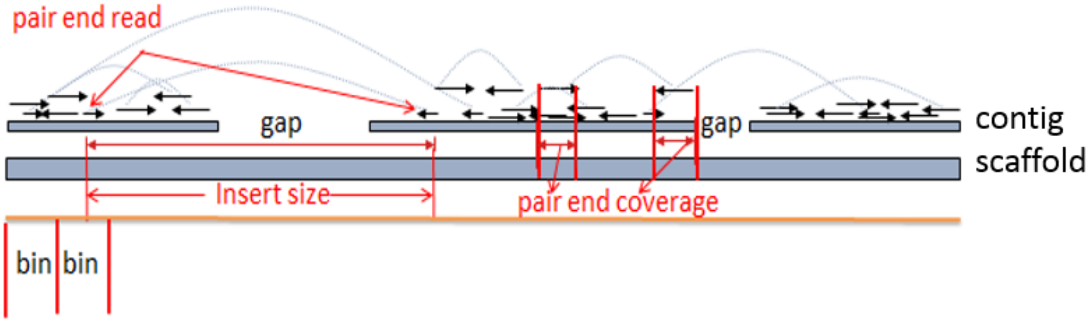
tstart	tend	qstart	qend	tmaplen	smaplen	%IDY	tlen	qlen	tcov	qcov	tname	qname
878024	878325	1	302	302	302	100	1531933	302	0.02	100	gi|330443520|ref|NC 001136.10	C16395
884200	892717	1	8447	279	280	99.64	1531933	40433	0.02	20.77	gi|330443520|ref|NC_001136.10	scaffold148
892711	909745	1	17035	17035	17035	100	1531933	17035	1.11	100	gi|330443520|ref|NC_001136.10	C19617
567739	568761	1023	1	1023	1023	100	1090940	1028	0.09	100	GI|330443578|ref|NC_001139.9	C17163
571977	573589	1613	1	1613	1613	99.88	1090940	1613	0.15	100	GI|330443578|ref|NC_001139.9	C17483
575307	593553	40433	20436	18045	18045	100	1090940	40433	0.6	44.62	GI|330443578|ref|NC_001139.9	scaffold148
581853	593353	13476	1976	11501	11501	100	1090940	13476	1.05	85.34	GI|330443578|ref|NC_001139.9	scaffold281
593363	611764	1	18402	18402	18402	100	1090940	18402	1.69	100	GI|330443578|ref|NC_001139.9	C19667
611748	627202	1	15455	15455	15455	100	1090940	15455	1.42	100	GI|330443578|ref|NC 001139.9	C19543

Consider the example scaffold 148 in Figure 3.2.b. It shows the alignment result between the scaffolds and the reference genome. From the alignment result, we can see that two different parts from this scaffold are located in two different chromosomes of the reference genome. One has 20.77% Identity, and the other one has 44.62% Identity. Both of them will be filtered in the final scaffold sets by the Identity criteria, as the default setting for ABACAS is 70%.

However, if it is a mis-assembled scaffold, it should be first split into two parts at the mis-assembly region. Then, for each part, the Identity parameter is computed based on the length of that part, not the total length of the scaffold. This way, both parts will be kept and put in different chromosomes in the final assembled genome.

In order to avoid subsequent assembly error when aligning to the reference genome, mis-assembled scaffolds are identified first in our work.

Figure 3.3 Model of assembly by pair end reads. The arrow denotes pair end reads



In a correctly assembled scaffold [like the example in Figure 3.3], there should be pair end reads spanning the whole region of this scaffold. In other words, each position in this scaffold should be covered by the insert size of some pair end reads. If some region is not covered this way, it may be mis-assembled. Based on this hypothesis, our method uses the insert size of pair end reads to identify mis-assembled scaffolds.

The process is detailed below:

- The region for each scaffold is divided into equal and contiguous bins (window) [bin=10bp] [Figure 3.3];
- After aligning the raw reads into the assembled scaffolds, the coverage of each bin is computed by the insert size of pair end reads. As a control, pair end reads that are put in bins that are too far apart or too close together---beyond the 95% confidence interval of the real insert size---are discarded.
- The zero-coverage regions are identified as potential mis-assembly regions.

Considering the same example shown in Figure 3.2.a, it shows the distribution of coverage of each bin across scaffold 148. From the result, we see that there is a region which has zero coverage. According to our hypothesis, this scaffold should be a mis-assembled scaffold.

After identifying mis-assembled scaffolds, we split them into several new scaffolds at the mis-assembly regions. Consider the example---scaffold 148---in Figure 3.2. There is just one mis-assembly region [between 845*10 and 2041*10]. Hence, we split it into two parts: 1-8450 and 20410-end.

3.2.3 Alignment to reference genome

After mis-assembly scaffold identification and correction, the new scaffold sets are aligned to the reference genome using MUMmer [101], due to its efficiency. MUMmer has commonly been used in the discovery of syntenic regions and chromosome-scale inversions [102]. For our purpose, we mainly use it for alignment between our corrected de novo assembly result and the reference genome. The alignment result helps us determine the order and orientation of scaffolds. However, as a quality control, any scaffold which has a low **Identity**[<80%] on the reference genome will be filtered.

3.2.4 Repeat scaffold identification

As we mentioned earlier, one major drawback of ABACAS [56] is that it just randomly assigns each repeat scaffold to a matching location in the reference genome. However, a real repeat scaffold should have multiple alignment locations in the reference genome, and we should retain all of them. For a non-repeat scaffold, if it has multiple alignment locations in the reference genome, this may be because of some rearrangement between the reference genome and the target genome; in this case we should retain the target location with the highest Identity and matching quality, not all locations.

Therefore, before dealing with the repeat scaffolds in the alignment result, we should first check which one is a real repeat scaffold. Repeat scaffold always refers to DNA sequences that are present in multiple copies in the genome in which they reside. Hence, for genomic sequencing, the reads abundance for sequences corresponding to these regions is always higher than other regions.

Based on this fact, our method for identifying repeat scaffolds is similar to mis-assembled scaffold identification. The only difference is that the windows/bins coverage is computed by pair end reads coverage [Figure 3.3], not by insert size. The pair end reads coverage means the read coverage for specific regions, which mainly measures the abundance for this scaffold.

It also contains three steps:

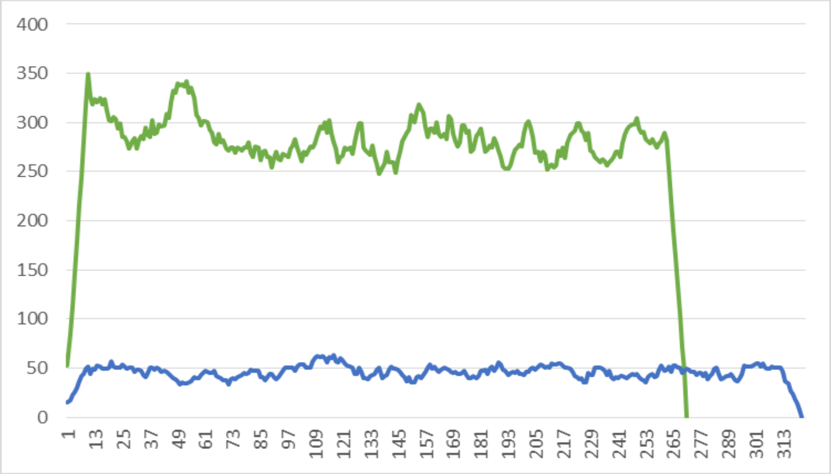
- The region for each scaffold is divided into equal and contiguous bin (window) size [10bp] [Figure 3.3].
- The coverage of each window is computed by the coverage region of each reads.
- Average coverage for each scaffold is given by:

$$average\ coverage = \frac{\sum coverage\ for\ each\ bin}{number\ of\ bin}$$

> Note: for a scaffold with gaps, the bin of each gap is excluded in the calculation—because no reads are mapped to the gap.

If the average coverage of a scaffold is much bigger than the expected value [the default value in our pipeline is mean+2*stdev], it may be a repeat scaffold [Figure 3.4].

Figure 3.4 An example coverage comparison between a repeat scaffold and a non-repeat scaffold



3.2.5 Overlap scaffold identification

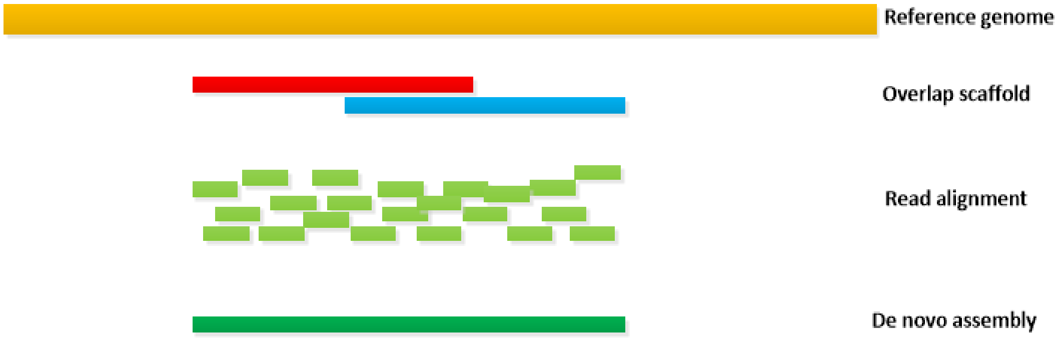
Another case that needs attention is that, there are some overlapping scaffolds after alignment to the reference genome. This may be due to some partial repeat region or tips of repeat sequence during de novo assembly, like the example shown in Figure 3.5. Before ordering and orientating the scaffolds, we also need to develop a method to deal with this type of scaffolds.

Our method for dealing with this case can be summarized into two steps; see Figure 3.5:

Step 1: The raw reads aligned to these regions are extracted to form a new read set.

Step 2: Using this new raw read set, we redo the de novo assembly for this region and obtain a unified scaffold.

Figure 3.5 A method to deal with the overlap scaffolds



In summary, combining these five components, we can order and orientate all the scaffolds located in the same chromosome of a reference genome into a much longer one. By this new method, we can considerably improve the final assembly result.

3.3 Results

3.3.1 Evaluation on gold-standard dataset

Before applying the proposed reference-based genome assembly method to our real plant sample, we first test each component of our method on a gold-standard data set [103] from a simulated genome [112,498,656bp] produced by the Evolver suite of genome evolution tools (http://www.drive5.com/evolver). Evolver can simulate the forward evolution of multi-chromosome haploid genomes, and it includes models for evolutionary constraint, protein codons, genes and mobile elements. We use it to generate several pair end datasets with different insert libraries, according to the error model that Illumina protocols introduce.

In the final generated sequencing libraries, the total coverage for the whole genome is around 120X [Table 3.3]. In order to test the effectiveness of our method on low-coverage sequencing results, we randomly selected several subsets of this (repeat 3 times at each coverage level in our study) dataset and finally get test sets with 100x, 80x, 60x, 40x, 20x and 10x coverage (3 replicates).

Table 3.3 Statistic of sequencing information for gold dataset

library	read num	seq length	total base	coverage	pair end
200bp	22,499,730	100	4,499,946,000	40X	yes
300bp	22,499,730	100	4,499,946,000	40X	yes
3000bp	11,249,867	100	2,249,973,400	20X	yes
10000bp	11,249,867	100	2,249,973,400	20X	yes

3.3.2 Evaluation of mis-assembly detection component

An important component in our pipeline is mis-assembly scaffold identification. Therefore, we evaluate its accuracy on the gold-standard data, which has been introduced in section 3.3.1.

First, we apply several known de novo assembly tools on the test sets at different coverage (120x, 100x, 80x, 60x, 40x, 20x, 10x) of the gold-standard dataset. Then, we use our mis-assembly component to evaluate them. The number of mis-assembled scaffolds reported for various de novo assembly methods is shown in Table 3.4.

At the same time, for this gold-standard dataset, it has a known genome. Hence, by aligning the assembled scaffolds to the reference genome, we know which scaffold is a real mis-assembled scaffold and where the mis-assembly regions are.

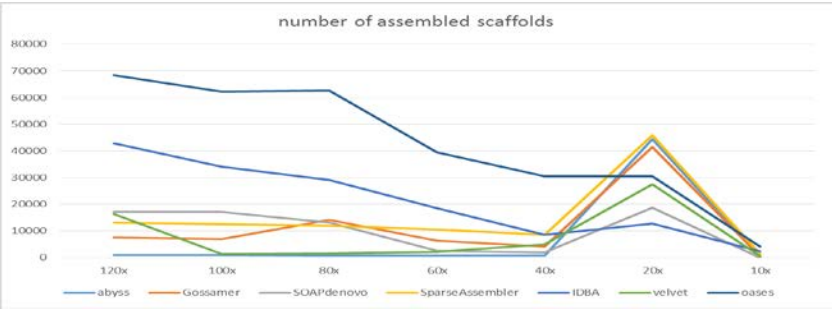
Considering the same example shown in Figure 3.2.b, after aligning scaffold148 to the reference genome, one part [1-8,447] and another part [20,436-40,433] are located in two different chromosomes. By our identification method shown in Figure 3.2.a, the mis-assembly region is also located in region spanned by the 845th to 2041th bins. After multiplying by bin size, this is [8,450-20,410]. Hence, the result produced by our method is consistent with the real mis-assembled region.

Table 3.4 shows the average number of scaffolds reported by our method to be mis-assembled when it is applied to the test sets at various levels of coverage. From the results, we find that there are fewer mis-assembled scaffolds at higher sequencing coverage. It may be because the accuracy for these de novo assembly methods are higher at higher sequencing depth, such as having more reads for resolving repeat bubbles in the de Bruijn graph. However, at 20x coverage, there is a huge increase of possible mis-assembled scaffolds. An explanation can be inferred from the total number of assembled scaffolds shown in Figure 3.6. Generally, the total number of assembled scaffolds is decreasing with decrease of sequencing coverage, because more and more regions (especially those of lower abundance) of the genome may not be included in the lower-coverage test sets. However, there is a big increase of assembled scaffolds at 20x coverage; this is because some previously covered regions (even those of higher abundance) get broken into short fragments in the final assembly at 20x coverage. At this low 20x coverage, it is much more difficult to deal with repeat bubbles, which in turn results in much more mis-assembled scaffolds. At 10x coverage, many of these fragmented higher abundance regions may be also not be included in the 10x coverage test set, leading to very few scaffolds.

Table 3.4 Mis-assembly result based on the gold-standard data from Assemblathon 1 [103]. **The number means the average number of mis-assembled scaffolds reported by our method.**

	120x	100x	80x	60x	40x	20x	10x
abyss	2	1	1	1	3	4,786	1
Gossamer	13	17	23	15	10	2,496	3
SOAPdenovo	35	56	20	12	4	2,177	0
SparseAssembler	1,436	1,458	1,408	1,398	1,219	21,687	184
IDBA	1	1	14	2	3	1,363	207
velvet	7	1	1	75	2	10,556	93
oases	12	3,879	4,327	5,090	5,631	4,069	5,479

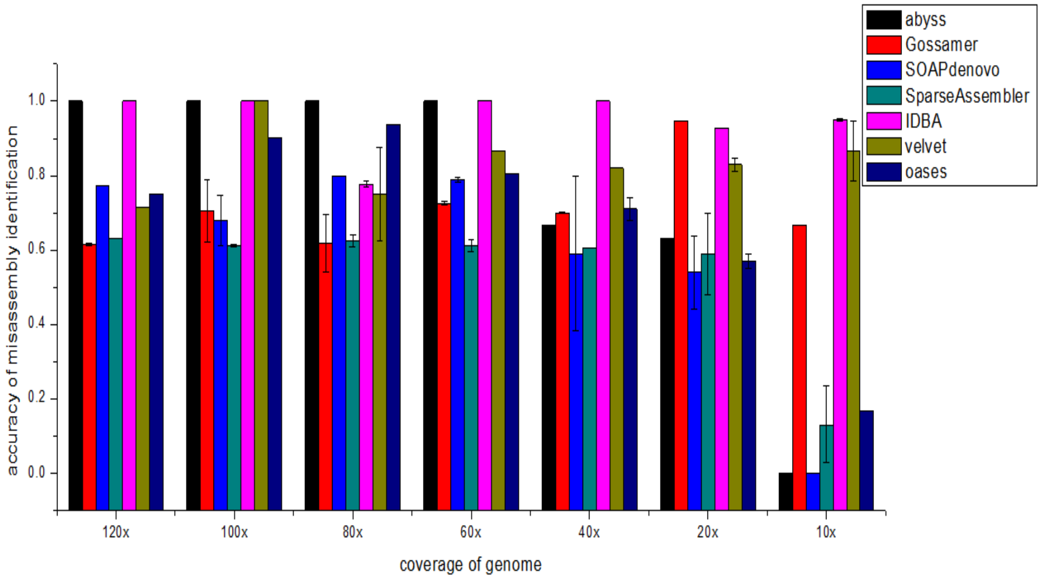
Figure 3.6 Average number of assembled scaffolds by different de novo assembly methods



As shown in Figure 3.7, at 80x coverage, about 80% of the reported mis-assembled scaffolds have been verified correct. At 60x coverage, 75% of them are correct. Even at 40x coverage, around 65% of them are correct. Only at very low coverage (20x and 10x), the percentage is a little low.

However, for our purpose, even around 65% accuracy is acceptable. This is because, if a scaffold was wrongly identified by our method to be mis-assembled, the segments (resulting from the split in our method) actually would align side by side in the reference genome. Consequently, these scaffolds would also be kept and not so much information is loss. On the other hand, the correctly identified mis-assembled scaffolds, if they were not identified and corrected, would likely be discarded as they would not match the reference genome at a sufficient level by Identity criteria, resulting in a loss of information.

Figure 3.7 Percentage of correct mis-assembled scaffolds reported by our method for each de novo assembly method under different coverage of the raw genome



In summary, based on the evaluation on different coverage of this gold-standard dataset, our method for mis-assembled scaffold identification has relatively high effectiveness.

3.3.3 Evaluation of repeat-scaffold detection component

Similarly, before applying the repeat scaffold identification method to our real plant samples, we also test it on these same gold-standard data sets [103] used earlier in mis-assembled scaffold identification.

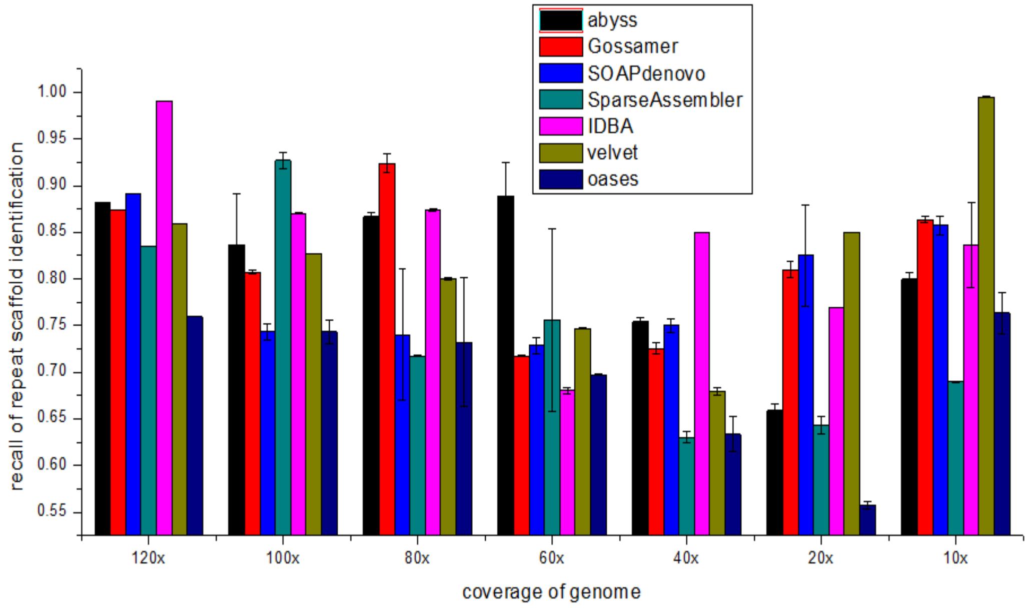
After aligning the assembled scaffolds to the known reference genome, the number of scaffolds having multiple locations in the reference genome is shown in Table 3.5.

Table 3.5 Repeat scaffold result based on the gold-standard data from Assemblathon 1 [103]. The number is the average number of scaffolds mapped to multiple locations in the reference genome for different methods.

	120x	100x	80x	60x	40x	20x	10x
abyss	442	539	353	535	126	293	30
Gossamer	2,563	2,098	6,614	1,082	91	315	22
SOAPdenovo	7,634	6,759	1,671	1,785	56	132	21
SparseAssembler	9,710	8,151	3,937	5,928	276	300	58
IDBA	31,586	21,455	3,505	3,846	207	234	144
velvet	61,586	41,763	29,658	9,782	4,793	686	638
oases	28,785	24,042	7,320	8,761	317	312	95

Using our repeat-scaffold identification component, more than 80% of these multi-location scaffolds can be detected, for the 120x, 100x, and 80x test sets [Figure 3.8]. Even the lower-coverage test sets, such as 60x and 40x, our method achieves around 75% recall [Figure 3.8].

Figure 3.8 Recall for our repeat scaffold identification component



However, as shown in Figure 3.9, the precision of our repeat-scaffold identification method is only around 50%. There may be several reasons for this problem:

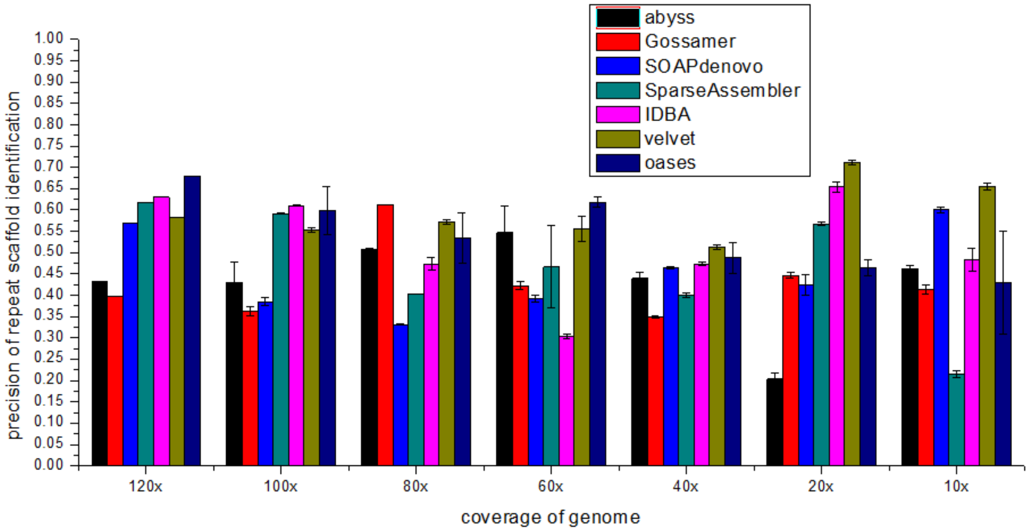
First, some of these may be real repeat scaffolds; however, they are missed in the reference genome due to sequencing error or other reasons. These false positives are potentially true positives; their status may be confirmed by checking against known repeat motifs.

Second, some of these may be in regions that are over amplified. It may be possible to eliminate this category of false positives by checking whether their flanking regions also have high abundance.

Third, there may be a divergence between the reference genome and the genome being assembled.

Although the precision is only about 50%, this is also not critical for our purpose. A scaffold is a false-positive repeat scaffold means that it does not align to multiple locations in the reference genome. Thus it is mapped to at most one location in the reference genome, and no information is lost. In contrast, if the recall is low, it would cause a loss of information, as a missed repeat scaffold that should be mapped to multiple locations in the reference genome is mapped to only one location.

Figure 3.9 Precision for our repeat scaffold identification component



Hence, based on our method, if a scaffold has multiple alignment locations in the reference genome and is identified as a repeat scaffold by our method, we retain all the locations and multiple copies in the final scaffold sets. However, for a scaffold that is not a repeat scaffold under our criteria, we just retain the copy with the highest match score, even when it has multiple alignment locations in the reference genome [Figure 3.1]. This may be because of rearrangement in the reference genome.

3.3.4 Evaluation of overlap-scaffold detection component

Similar to the previous two sections, we also investigate the statistics of overlap scaffolds in the gold-standard data [103]. From Table 3.6, it is clear that overlap scaffold groups form a relatively high portion even for this small simulated genome. It is obviously necessary to redo the assembly for overlapping scaffold groups in real datasets.

Table 3.6 Average number of overlap scaffold groups based on the gold-standard data from Assemblathon 1 [103] at different coverage.

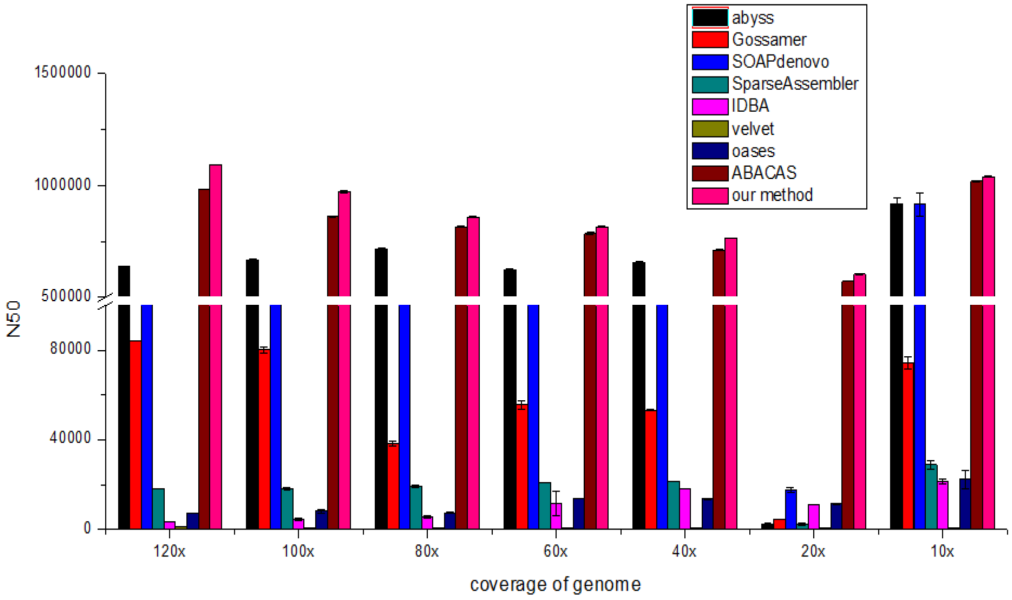
	120x	100x	80x	60x	40x	20x	10x
abyss	1,426	1,380	1,238	958	884	7,433	98
Gossamer	1,468	1,405	1,429	342	155	16,202	23
SOAPdenovo	16,663	17,058	8,143	2,267	402	5,064	8
SparseAssembler	7,178	6,259	4,484	2,829	1,976	5,089	292
IDBA	40,635	33,672	18,192	17,207	3,066	4,458	552
velvet	66,688	56,630	79,159	87,799	73,521	55,701	16,391
oases	41,878	38,159	26,990	12,282	8,833	9,873	1,028

In summary, by these evaluations, we find that these important components in our pipeline are necessary and have relatively high accuracy. In the next Chapter, we will discuss how to apply our reference-based genome assembly pipeline in a real plant genome project, and make comparison with de novo assembly tools and other reference-based genome assembly tools.

3.3.5 Comparison between de-novo and reference-based genome assembly

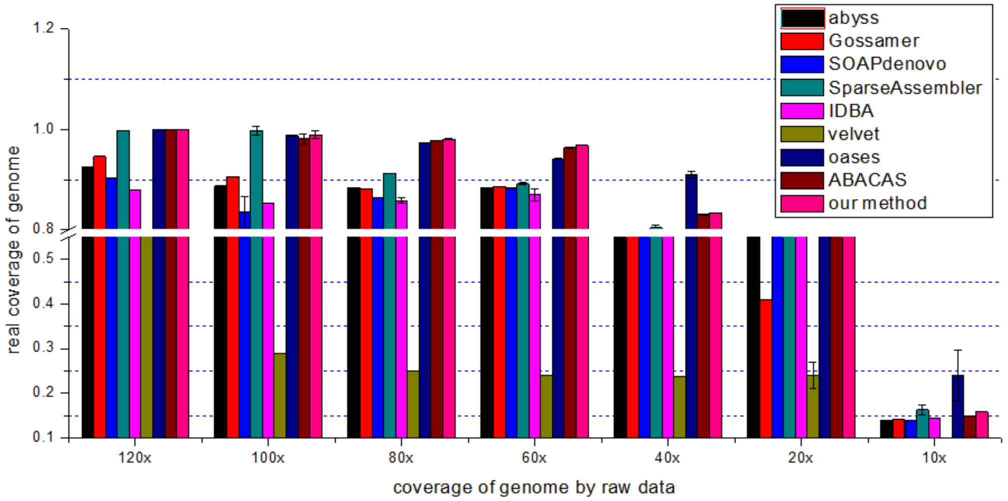
In order to better appreciate the performance of de-novo assembly methods and reference-based assembly methods, we also compare the final results on the same gold-standard dataset between these different groups of methods [Figure 3.10 and Figure 3.11].

Figure 3.10 N50 for different method under different coverage of genome.



From the result shown in Figure 3.11, we can see that if we want to get more than 90% coverage of the whole test genome using de novo assembly, the raw sequencing coverage has to be at least 100x, even 120x. With 80x and 60x raw sequencing genome coverage, we can just have around 85% coverage for genome. However, by reference-based genome assembly, even at 60x sequencing coverage, we can get around 90% of genome coverage. From the N50 distribution in Figure 3.10, reference-based genome assembly methods also outperform de novo assembly methods at all coverage.

Figure 3.11 Final genome coverage by de novo assembly methods. Genome coverage=total number of bases of final scaffolds/genome size



Hence, from these comparisons on final assembly results, we see that reference-based genome assembly [ABACAS and our method] methods outperform de-novo assembly methods, not only on N50, but also on whole-genome coverage [Figure 3.10 and 3.11]. In addition, our method slightly outperforms other reference-based methods, such as ABACAS, on this gold dataset.

3.4 Conclusions

In this chapter, we have proposed a new reference-based genome assembly pipeline. The main novel features for our pipeline are the techniques for detecting and handling mis-assembled scaffolds, repeat scaffolds and overlap-scaffolds.

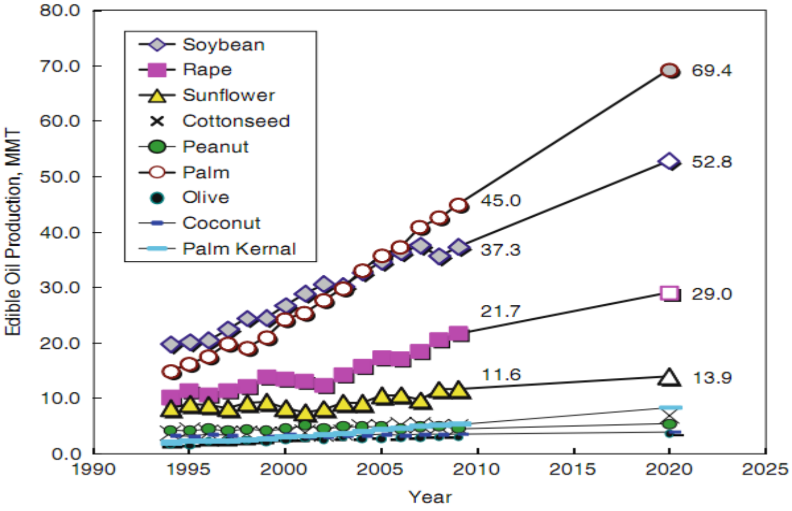
From the evaluation on a gold-standard dataset, we find that these major novel components in our pipeline have relatively high accuracy.

Chapter 4

APPLICATION ON OIL PALM

Lipids, a major class of primary metabolites, also called fat/oil at room temperature, are an essential part of the human diet. Many plant seeds accumulate storage products during seed development to provide nutrients and energy for seed germination and seedling development. Some seed crops---such as corn, wheat, rice, peas and common beans---accumulate starch as the main form of energy storage in the seeds. However, oilseeds---such as soybean, corn, coconut, jatropha and oil palm---accumulate oil instead of starch. Together, these oilseed crops account for 75% of the world vegetable oil production. These oils are used in the preparation of many kinds of food, both for retail sales and in the restaurant industry. Among these main oil crops, oil palm is the highest oil-producing crop in the world's oil market since 2005 [Figure 4.1].

Figure 4.1 Trends in global production of major plant oils [1]



4.1 Background

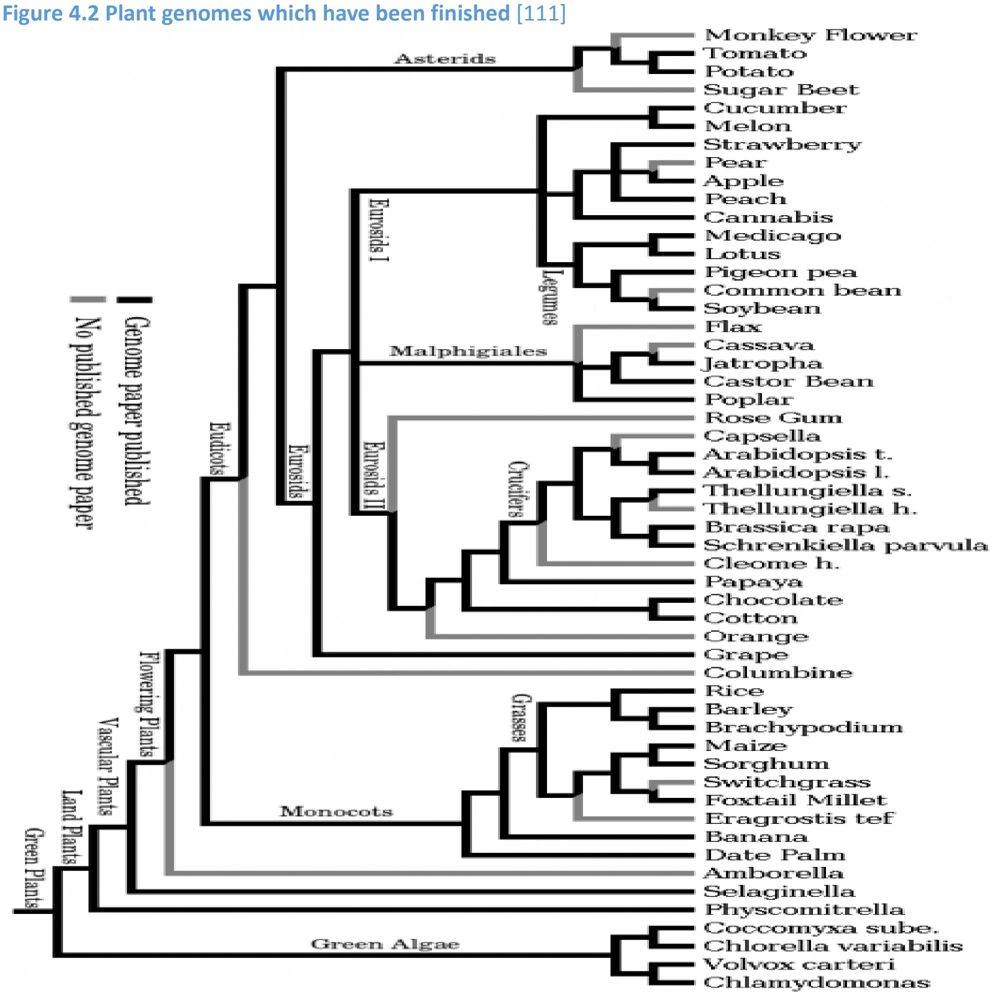
To increase the production of oil crops, a simple method is to increase the planting area. However, it is not sustainable to keep extending planting area, because of increased competition for land by the rapidly rising population. Therefore, it is a better strategy to increase fruit/seed oil content than to increase the planting area.

In recent years, molecular genetics approaches, based on homolog search or screening, have been successfully used to modify seed oil content for several plants. For example, over-expression of a diacylglycerol acyltransferase (AtDGAT) cDNA in wild-type *Arabidopsis thaliana* enhanced oil deposition and average seed weight [8]. The research of Wang et al. suggests that oil content of soybean seeds can be increased by up-regulation of two soybean Dof-type transcription factor (GmDof) genes, that are associated with fatty acid biosynthesis [9]. Several researchers have found that the mutation of FAD2 and FAD3 can regulate oil composition and elevate oleic acid levels [104-107]. Although the biochemical pathways that produce different oil components are well characterized, there is still no genome-wide model to identify new genes/enzymes involved in lipid biosynthesis.

With the development of high-throughput technologies, including the newly developed Solexa/Illumina RNA-sequencing, new genes/specific transcripts can be easily discovered and analyzed on a genome-wide model. In the research of Severin et al. [15], RNA-seq provided a record of high-resolution gene expression in a set of various tissues for soybean. By differential analysis, they also found dramatic highly-expressed genes and the genes specific to legumes in seed development and nodule tissues. Different from their RNA-level analysis, based on genomic sequencing techniques, the whole genome for many oil crops such as soybean [2], corn [108], sesame[109], coconuts [110] and jatropha [10] have been dissected [111]; see Figure 4.2. Thus, many new and specific lipid biosynthesis genes/enzymes, and even new biosynthesis pathways, have been discovered using these draft genomes [105, 107, 112].

However, as the highest oil-yielding crop, whole-genome sequence and molecular resources available for oil palm still remain scarce [113]. Only several months ago, while this work was in progress, a draft genome of Pisifera was released [18, 114], which is different from the commercial Dura strain we work on. To provide oil palm researchers with additional resources to study and improve this important oil crop, we attempt to construct a draft genome and transcriptome resources for Dura based on next-generation sequencing data sets.

Figure 4.2 Plant genomes which have been finished [111]



4.2 Methods

4.2.1 Whole-genome short-gun (WGS) sequencing for oil palm

The increasing use of next-generation sequencing (NGS) has resulted in an increased growth of the number of de novo assembled genomes [Figure 4.2]. In our study, short-insert pair-end (clone size: 300 bp) and large-insert mate-pair (3-5,10 and 20 kb) libraries were prepared and sequenced by Illumina and 454 technologies following the manufacturers' instructions, and the resulting sequences were used to assemble the Dura draft genome [Table 4.1]. Summarizing these sequence data, we obtained 92X sequence coverage for the entire genome.

Table 4.1 Sequence library for Dura by next-generation sequencing platform

insert size	avg size (bp)	raw reads	usable reads	usable base (bp)	Pair end	depth	Technology
300	101	558,695,836	406,267,011	39,396,057,965	Yes	43.8	Illumina
300	101	374,109,317	325,921,739	32,366,281,947	Yes	36	Illumina
3-5k	51	144,845,306	144,845,306	7,387,110,606	Yes	8.2	Illumina
3-5k	51	37,083,563	37,083,563	1,854,178,150	Yes	2.1	Illumina
10k	51	26,355,787	26,355,787	1,344,145,137	Yes	1.5	Illumina
20k	404.08	558,411	558,411	225,642,316	yes	0.13	454
NA	378.28	1,322,072	1,322,072	500,110,270	no	0.28	454
Total							
		1,142,970,292	942,353,889	83,073,526,391		92	

4.2.2 Reference-based genome assembly

Considering the large genome size of our Dura sample and the released genome of another strain of oil palm, our proposed reference-based genome assembly is adopted to construct the draft genome for Dura oil palm [Figure 3.1].

In the de novo assembly part, SOAPdenovo [55], ABySS [50], IDBA [54], Velvet [48], Oases [52], Gossamer [51], SparseAssembler [53] and Allpaths-LG [49] are selected for comparison.

4.3 Results

4.3.1 Evaluation method

Several metrics are commonly used for assessing the assembly results.

- Number of the assembled scaffolds
- Total length of the assembled scaffolds
- Length of the largest scaffold

- N50 of contigs/scaffolds: N50 is similar to a mean or median, but greater weight is given to the longer scaffolds. Given a set of scaffolds, each with its own length, the N50 length is defined as the length for which the collection of all scaffolds of that length or longer contains at least half of the total of the lengths of the scaffolds. Sometimes, some researchers also show the N90, N20 value in the final comparison.
- Percentage of the gaps in the final scaffolds. During the de novo assembly, many gap regions are introduced into the scaffold sets [Figure 3.3], due to use of long insert library.

Hence, in the following comparison, we mainly focus on these metrics.

4.3.2 Comparison between de novo assembly and reference-based assembly

Based on the metrics in section 4.3.1, we compare here our proposed method with de novo assembly methods.

Comparing the assembly results between several stand-alone de novo assembly tools---viz, SOAPdenovo [55], Velvet [48], IDBA [54], oases [52], SparseAssembler [53], Gossamer [51], ABySS [50]---and several tools which just do scaffolding--- viz, Opera [115], SSPACE [116], SOPRA [117]---it is clear that SOAPdenovo outperforms other methods on the final scaffold level, especially on N50 [Table 4.2 and 4.3]. Hence, for the de novo assembly component in our pipeline, due to the larger N50 and longer largest scaffold, SOAPdenovo is adopted in this step. However, even for the best results among these several tools, N50 is only around 13,000 bp, which means there are still many short scaffolds. Based on de novo assembly methods, if we want to improve the assembly results, more insert libraries, especially large insert libraries are needed.

In order to reduce the cost and take advantage of the released genome sequence of another strain of oil palm, we adopt our reference-based genome assembly pipeline to improve the results, as described in section 3.2.

Table 4.2 Comparison between different de novo assembly tools on Contig level

	Contig level			
method	number	largest	N50	total bases
SOAPdenovo	31,043,382	60,406	91	2,742,444,815
AByss	21,671,156	93,899	103	2,192,693,039
IDBA	1,021,976	51,470	1,582	747,411,128
Sparseassembler	22,378,344	33,450	88	1,830,152,485
Velvet	35,672,351	50,472	86	2,821,064,376
Gossamer	No contig level result			
Allpaths-LG	Not enough memory			

Table 4.3 Comparison between de novo assembly methods and our proposed reference-based method

	Scaffold level			
	number	largest	N50	total bases
SOAPdenovo	1,026,189	270,947	13,984	1,556,659,866
SOAP+Opera	30,924,876	114,186	91	2,749,023,079
SOAP+SSPACE	30,338,306	165,586	91	2,741,120,819
SOAP+SOPRA	28,765,874	184,764	158	267,287,543
AByss	21,671,129	93,899	2,030	2,192,705,185
AByss+SSPACE	21,487,200	135,159	103	2,196,888,385
Abyss+Opera	21,671,156	93,899	103	2,192,693,039
IDBA	707,194	85,211	3,898	726,160,083
IDBA+SSPACE	685,306	255,288	3,777	744,070,911
IDBA+Opera	652,984	327,876	4,239	752,187,646
IDBA+SOPRA	679,268	297,982	4,031	748,982,674
SparseAssembler	22,378,344	33,450	88	1,830,152,485
Sparse+SSPACE	21,960,979	292,071	88	1,831,195,921
Sparse+Opera	22,678,344	33,874	88	1,836,172,673
Sparse+SOPRA	21,987,372	34,127	86	1,854,092,132
Velvet+Oases	11,329,281	135,396	785	1,873,254,194
Gossamer	20,107,482	127,936	2,673	2,046,871,965
Allpaths-LG	Not enough memory			
Our method	608,380	22,365,697	576,146	2,584,445,363

Comparing our reference-based assembly results with de novo assembly methods, we see that not only N50, but also the longest scaffold, are improved a lot; c.f. Table 4.3, which shows the specific advantage of our reference-based genome assembly pipeline.

4.3.3 Comparison between ABACAS and our proposed method

In order to explore the effect of each major component in our final results, we also try to compare the result between our proposed reference-based genome assembly pipeline with a popular reference-based genome assembly method—ABACAS.

Comparing with ABACAS, first, the number of scaffolds is reduced by around 170,000 [Table 4.4]. Another important improvement is that the number of scaffolds which can be mapped to the reference genome is improved a lot [Table 4.4]. This part has largely benefitted from several core components---namely, mis-assembled scaffold identification, repeat-scaffold identification and overlap-scaffold identification---in our pipeline, which we will explain in detail in following section.

Table 4.4 Comparison between ABACAS and our method

	ABACAS	Our proposed method
#scaffold	775,109	608,380
largest scaffold	22,002,004	22,365,697
Scaffolds located in reference genome	264,600	594,782
N50	501,301	576,146
total	2,646,425,608	2,584,445,363

4.3.3.1 Effect of mis-assembly identification component

Using our mis-assembled scaffold identification method, around 28,585 mis-assembled scaffolds have been identified [Table 4.5]. After our correction and re-alignment to the reference genome, around 26,118 (91.2%) can be located in the reference genome again, most of which will be likely incorrectly filtered by ABACAS. If these mis-assembled scaffolds were not identified and corrected, fewer scaffolds would be connected by the reference genome.

Table 4.5 Mis-assembly information in our pipeline

	SOAPdenovo
#scaffold	1,026,189
mis-assembly	28,585
# scaffold located in reference genome for mis-assembly set	26,118

Therefore, from this comparison, it is important to deal with mis-assembled scaffolds in reference-based genome assembly methods, because a lot of mis-assembled regions are introduced. Otherwise, most of them will be filtered due to lower Identity.

4.3.3.2 Effect of the repeat-scaffold identification component

After aligning the de novo assembly scaffolds to the reference genome, 45,902 scaffolds have multiple match locations [Table 4.6], and a total of 127,195 match locations in reference genome. Among these 45,902 repeat scaffolds, 27,900 (61%) are predicted as potential repeat scaffolds by our method.

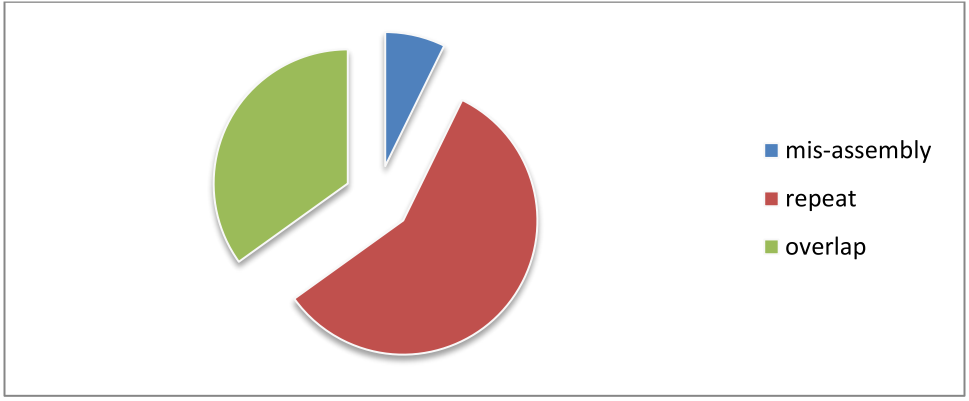
Table 4.6 Statistic for the repeat scaffolds

#repeat scaffold located in reference genome	45,902
#total repeat times	127,195
#potential repeat scaffold by our threshold	27,900

Considering the huge number of repeat scaffolds, it is not reasonable to randomly assign a match location for these repeat scaffolds, as per ABACAS. Hence, it is necessary to give a reasonable location for these repeat scaffolds.

In summary, we have mentioned in the beginning of this section that the number of scaffolds is reduced by around 170,000, and the scaffolds that can be located in the reference genome is increased by 330,182, in comparison to ABACAS. Among these 330,182 scaffolds, three major components account for this improvement. One is the mis-assembly component, which we have explained in section 3.2.4.2. Another one is the repeat-scaffold component, which has been shown in section 3.2.4.4. The other reason is because of overlap scaffold, which has been shown in section 3.2.4.5. The percentage distribution is shown in Figure 4.3.

Figure 4.3 Pie chart of the increased scaffold located in reference genome, comparing to ABACAS



From the comparison between ABACAS and our proposed pipeline, the major components in our pipeline have great effect on the final results.

Benefitting from next-generation sequencing, draft genomes for many species have been finished [2, 118, 119]. Afterward, post-processing and revising these draft genomes become a challenge. Generally, a little improvement in the assembly can save a lot of time and cost in the post-processing step. Therefore, we believe our refined reference-based genome assembly pipeline can provide some evidence and guidance for the further improving reference-based genome assembly methods.

Due to the advantages for our reference-based genome assembly pipeline, we perform downstream analysis using the assembled Dura draft genome produced by our proposed pipeline.

4.4 Evaluation of Dura draft genome

Before using the Dura draft genome for the downstream analysis, we should first evaluate its quality. If the quality is not acceptable, it can cause many errors in the downstream analysis. At the same time, we can also use these evaluations to compare the accuracy of ABACAS with our reference-based genome assembly pipeline.

Three methods--- viz, EST coverage, completeness of genome and linkage map---are applied to check the quality, as explained individually in the following sections.

4.4.1 EST coverage

A total of 41,695 oil palm expressed sequence tags (ESTs), collected from leaf and mesocarp tissues [120], were used to assess the gene coverage of this draft genome for oil palm. ESTs were aligned to the genome by BLAT [121], which can handle introns in DNA/RNA alignment. Only ESTs with alignment of identity>0.9 were retained.

Our result indicates that the draft genome has a high coverage of protein-coding genome regions (~80%); see Table 4.7. In other words, most of the ESTs/cDNAs have sequences represented in our draft genome. In addition, we also applied the same EST dataset to the ABACAS result, which only achieved around 69.39% coverage.

Table 4.7 Statistic result for the EST coverage of the Dura draft genome

Dataset	EST reads	Number	match number	covered by assembly
mesocarp	>500	1,126	848	75.31%
	>100	6,514	4,878	74.88%
	>50	10,251	7,782	75.91%
	>10	20,972	16,415	78.27%
	all	33,841	27,034	79.89%
leaf	>500	7	5	71.43%
	>100	87	54	62.07%
	>50	216	144	66.67%
	>10	1,893	1,438	75.96%
	all	7,854	6,340	80.72%
total		41,695	33,374	80.04%

4.4.2 Completeness of draft genome

To check the completeness of our draft genome, a computational method CEGMA [122] was adopted, which defined a set of very conserved protein families that occur in a wide range of eukaryotes. By the conserved proteins defined therein, it can measure the completeness of each genome.

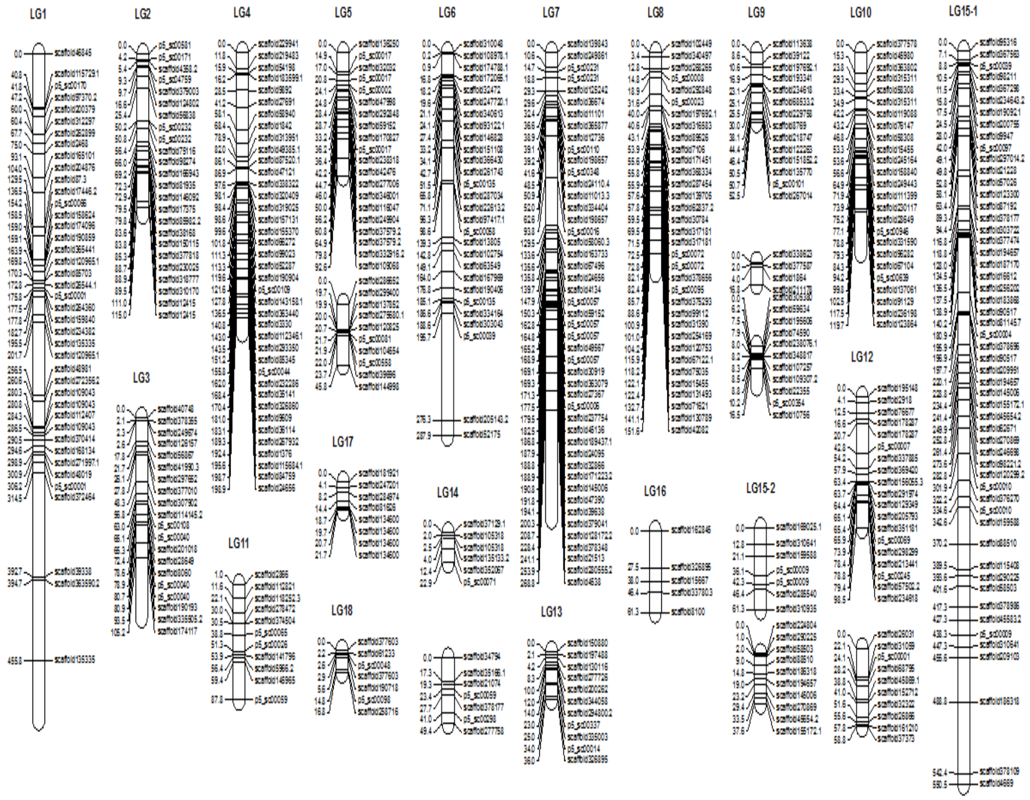
Among the 248 highly conserved proteins defined in CEGMA [122], 87% of them can be found in our draft genome. In other words, the result suggests that our draft genome uncovers ~87% of oil palm genes. However, for ABACAS, it only uncovers ~76% of oil palm genes.

4.4.3 Linkage map

Another method to evaluate quality is to use the marker dataset from oil palm and examine how many known markers can be found in our draft genome. Usually, researchers used markers for germplasm diversity analysis, linkage to monogenic traits of fruit color, shell thickness, and fatty acid composition of the oil [123, 124]. In our study, the linkage map can also help us to determine the correctness of each scaffold and map the scaffolds into real chromosomes.

We have already constructed an integrated linkage map consisting of 256 SSR markers from Billotte et al. [125] and 454 SSR markers identified by ourselves [Ref under review]. These 710 SSR markers were aligned to the draft genome using BWA [126] with no more than 1 mismatch. 98.03% of total markers can be successfully aligned to our draft genome [Figure 4.4], which is higher than ABACAS, with around 91.97% coverage. To some extent, this shows the high quality of our draft genome, which is consistent with the measurement in previous sections.

Figure 4.4 Relationship between linkage map and scaffolds in the draft genome of oil palm



Taken together, the results by three independent methods show the quality and completeness of our Dura draft genome. Therefore, this draft genome can be used for downstream analysis. In addition, the results also shows that our pipeline outperforms the popular reference-based assembly genome method—ABACAS.

4.5 Annotation of Dura draft genome

After assembling the Dura draft genome, the next task is annotation of this whole draft genome, such as protein-coding gene annotation, repeat annotation and noncoding RNA annotation. Without this information, cloning of some specific genes and connection between phenotype and genes is difficult. By identifying specific gene locus, it can be possible to connect gene, phenotype and function. Hence, in this section, we mainly discuss the different annotation for this draft genome.

4.5.1 Repeat annotation

For repeat annotation of our draft genome, it mainly contains three sources: de novo repeat finding, known repeat searching against existing databases and tandem repeat searching.

4.5.1.1 De novo identification of repeat sequence

RepeatScout [127] was used as the first step in de novo identification of repeat sequence in the draft genome. LTR retrotransposons were identified with LTR_FINDER [128] with default parameters. All repeat sequences with lengths >100bp and gap “N” less than 5% constituted the raw transposable (TE) library. Then, an all-versus-all BLASTN (E-value<1e-10) was used to search against the raw transposable element (TE) library, and a sequence was filtered when two repeats were aligned with identity >80% and minimal matching length >100bp. At the end, a non-redundant TE library was produced.

4.5.1.2 Identification of known TEs

RepeatMasker (version 4.0.1) (http://www.repeatmasker.org/) and the Repbase [129] database were used to find TE repeats in the assembled genome. TEs were identified both at the DNA and protein level. RepeatMasker was applied for DNA-level identification and RepeatProteinMasker was used to perform protein-level identification. Overlapping TEs were integrated to generate the final known TEs library.

4.5.1.3 Tandem repeats

Another important repeat family is tandem repeats. Hence, tandem repeats were also identified here using TRF [130], with parameters set to "Match=2, Mismatch=7, Delta=7, PM=80, PI=10, Minscore=50, and MaxPeriod=12". The same parameters have also been used in other organisms, such as the panda genome [97].

Finally, after combining three repeat libraries, repeat sequences account for 30.28% of the draft genome, similar to *Oryza sativa* (40%) [118] and *Brachypodium* (28%) [131], but much lower than *Zea mays* (84%) [108]. Same observation with the *Oryza sativa* [118], Sorghum bicolor [37] and *Zea mays* [108] genomes, the most abundant repeats in oil palm are retroelements [Table 4.8].

Table 4.8 Repeat statistics for oil palm draft genome

class	#	len(bp)	ratio
Retroelements	129,593	103,475,488	9.80%
DNA transposons	12,922	6,377,751	4.43%
Unclassified	1,039	385,351	1.02%
Satellites	260	62,938	0.00%
Simple repeats	174,807	12,960,688	0.89%
Low complexity	891,607	49,884,865	3.40%
total	1,158,469	163,569,879	19.54%
Tandem repeat	146,795	10,473,251	0.74%
unassembled repeats		173,574,126	10.00%
total repeats		563,569,879	30.28%

4.5.2 Gene annotation

4.5.2.1 De novo gene prediction

In order to improve the gene annotation results, several de novo prediction software programs---Augustus [68, 132] with gene model parameters trained on *Oryza sativa* , SNAP [133] and GeneMark-ES [134] with *Oryza sativa* parameter files---were used in our study. However, in the final gene sets, partial genes and small genes with sequence less than 200bp in length were filtered. These would contain much a higher error rate [119].

4.5.2.2 Evidence-based gene prediction

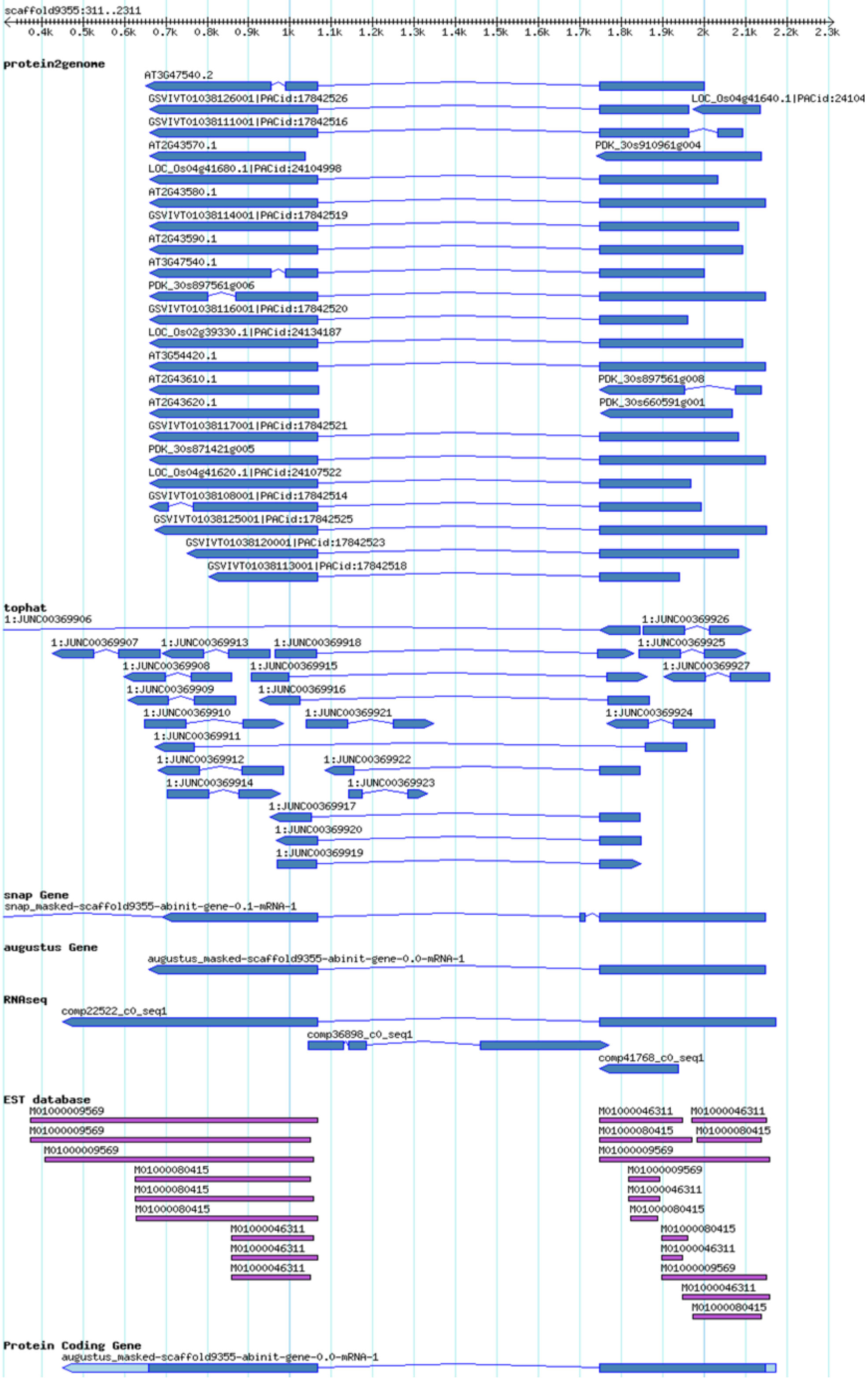
To provide more evidence complementary to the de novo annotation, we also attempted to integrate other evidences, like protein sequences, EST sequences and sequences derived from RNA seq.

(a) Protein sequence. Protein sequences from *Arabidopsis thaliana* [135], *Oryza sativa* [118], *Vitis vinifera* [35], and date palm [119] are used to provide protein domain evidence in our draft genome. The alignment between amino acid sequences of protein sequence and those in the draft genome was conducted using Exonerate [136].

(b) EST evidence. Totally, 41,695 EST/cDNA mesocarp and endosperm sequences of oil palm and 37,048 mesocarp EST sequences of date palm [120] are used to provide protein-coding evidence for the whole draft genome and gene families. This serves as the direct evidence for the protein-coding ability of the annotated genes.

(c) RNA seq evidence. Next-generation sequencing techniques have great potential to improve annotation quality, owing to their deep coverage and high throughput. Here, we collect 24 RNA sequencing samples from mesocarp [1.5, 2.5, 3.5, 4.5, 5.5 months] and endosperm [1.5, 2.5, 3.5 months] tissues in different development time points. For this RNA sequence dataset, Trinity [69] is first used to obtain the unique transcripts, which are then aligned to the draft genome to annotate the intron-exon structure for the annotated genes.

Figure 4.5 An overview of the gene prediction results by MAKER2 [126], visualized based on our developed database [137]



4.5.2.3 Reference gene set

Finally, the evidence-based and de novo gene sets are merged to form a comprehensive and non-redundant reference gene set by MAKER2 [138]. The final result is presented in Figure 4.5, which shows that the final annotated genome, and has captured the features both from de novo prediction and various evidence.

Based on this final gene set, we perform a general comparative analysis between the genes of oil palm and genes identified from *Arabidopsis*, *Sorghum*, rice and maize. Oil palm exhibits a high similarity to other species in parameters, such as the distribution of gene length, coding sequences (CDS), exon length and intron length; c.f. Table 4.9. Of the compared species, only the dicot *Arabidopsis* is obviously different from the other species with respect to gene length and intron length. This may be also the difference between dicot and monocot plants.

Table 4.9 Comparison of oil palm with other plants on gene number, average exon/intron length and other parameters. Gene density: the number of gene per 10kb

	sorghum	maize	rice	Arabidopsis	oil palm
gene number	27,640	32,540	34,792	25,498	36,015
average length	2,873	3,757	3,039	2,011	3,573
gene density	24	13.1	11	4.5	49.9
average exon per gene	4.7	5.3	3.7	5.2	3.8
average exon len	268	304	256	250	217
average intron len	436	516	409	168	522

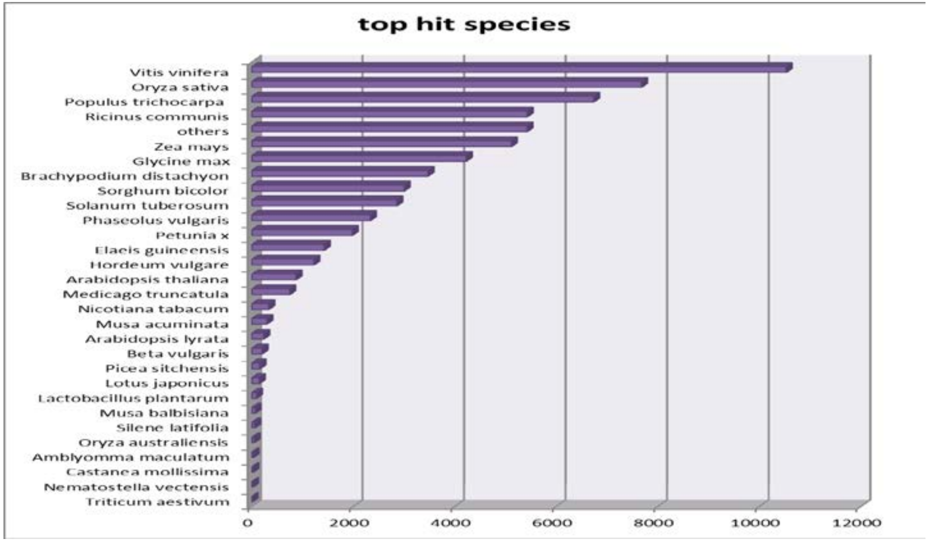
4.5.2.4 Gene function annotation

For the final reference gene set, function annotation is also necessary for biological research. Commonly, function annotation is assigned by homology search in other species.

In our study, the function for each gene was assigned using BLAST2GO [139], which attempted to find significant BLAST similarity to proteins from other organisms in the non-redundant (NR) database at the National Center for Biotechnology Information (NCBI). However, in the NR database, a lot of proteins are annotated with “predicted protein” or “conserved hypothetical protein”. Hence, in order to improve the annotation quality and reduce the effect of these missing annotation in the NR database, homology search using *Arabidopsis thaliana*, *Oryza sativa*, and *Vitis vinifera* protein databases is also provided for our reference gene sets.

The species which has the highest homology with our Dura sample is *Vitis vinifera* [Figure 4.6], a eudicotyledonous crop, followed by another monocotyledonous crop *Oryza sativa*. This high protein sequence similarity between the two less phylogenetically related plants (the monocotyledonous oil palm and eudicotyledonous grapevine) has been also observed by others, such as the date palm [119] and oil palm ESTs [140]. One possible reason may be owing to the completeness of *Vitis* genome and the higher gene number for *Vitis*. To explain this observation and detailed mechanism, additional studies are required.

Figure 4.6 The number of homologous genes in each species



4.5.3 NcRNA annotation

Noncoding RNAs (ncRNAs) are transcripts that are not translated to proteins but act as functional RNAs. Several well-known ncRNAs such as transfer RNAs (tRNAs) or ribosomal RNAs (rRNAs) can be found throughout the tree of life. Fulfilling central functions in the cell, these ncRNAs have been studied for a long time [141].

However, over the past years, a few key discoveries have shown that ncRNAs have a much richer functional spectrum than anticipated. The discovery of microRNAs (miRNAs) and short interfering RNAs (siRNAs) change our view of how genes are regulated. They play important roles in biological systems of eukaryotes by suppressing expression of target genes at the transcriptional and/or post-transcriptional level. Another surprising observation revealed by high-throughput methods is that, in human, 90% of the genome is transcribed at some time in some tissues.

Hence, besides the gene annotation for the draft genome, various ncRNA annotations are also performed in our study.

4.5.3.1 Identification of tRNAs

A transfer RNA (tRNA) is an adaptor molecule composed of RNA, typically 73 to 94 nucleotides in length, that serves as the physical link between the nucleotide sequence of nucleic acids (DNA and RNA) and the amino acid sequence of proteins. Knowing the tRNA repertoire of an organism is important because it affects the codon bias seen in highly expressed protein-coding genes.

Based on homolog search and secondary structure restriction, several tRNA identification software have been developed [142-144]. Among them, tRNAScan-SE [145] with default parameters has been successfully applied to predict tRNA genes in *Arabidopsis*, sorghum, maize, rice and date palm genome. Hence, we also use the same tool on our oil palm draft genome. Finally, our oil palm sample has a total of 622 predicted tRNAs, similar to 699 for *Arabidopsis* [146] and 606 for sorghum, suggesting that most of the oil palm tRNAs have been found; c.f. Table 4.10. It is interesting that 1 tRNA for Selenocysteine is detected in the oil palm genome, which has been only found in maize, sorghum and bamboo [96], but not in *Oryza sativa*, *Arabidopsis thaliana* and even the nearest species, date palm. The specific function for this tRNA needs further investigation.

Table 4.10 Compare oil palm with other plants on different class of tRNAs

	Z.ma	O.sa	S.bi	P.he	A.th	Date.P	Oil.P
tRNAs decoding Standard 20AA	1,413	720	535	1,076	685	399	571
Selenocysteine tRNAs (TCA)	4	0	1	6	0	0	1
Possible suppressor tRNAs	7	0	1	1	0	0	0

(CTA,TTA)							
tRNAs with undetermined/unknown isotype	14	0	8	2	1	3	1
Predicted pseudogenes	768	26	61	82	13	43	49
total tRNAs	2,206	746	606	1,167	699	445	622

4.5.3.2 Identification of rRNAs

Ribosomes are the molecular machines which form the connection between nucleic acids and proteins in all living organisms. The ribosomes dependence on ribosomal RNAs (rRNAs) for their function has caused them to be conserved at both the sequence and the structure level. Because of this, rRNAs are often used in comparative studies such as phylogenetic inference.

Commonly, rRNAs are often located by sequence similarity searches such as BLAST, due to the high level of sequence conservation in the core regions of the rRNA. The validity of the search results depends on the program and database used. Hence, we attempt to extract all the rRNAs from the Rfam database, which is the most comprehensive database for ncRNA. The rRNA fragments are identified by aligning the rRNA template sequences (Rfam [147] database, release 11.0) of *Arabidopsis thaliana*, *Oryza Sativa*, *Sorghum bicolor*, *Zea mays*, *Vitis vinifera* and several other plants using BLASTN with E-value at le-10, and identity cutoff at 90% or more. From the results shown in Table 4.11, we see that 1,182 rRNAs are found in the draft genome.

Table 4.11 Overview information of ncRNAs on oil palm draft genome

type	#	average length	total length	verification	% of genome
tRNA	636	76	47,240	303	0.00262%
rRNA	1,182	166	195,316	521	0.00465%
SnoRNA				164	

C/D box	139	169.5	23,555		0.0246%
H/ACA	124	102.5	12,710		0.0002%
snRNA	262	102.67	26,899	47	0.0009%
Known miRNA	199	127	181,979	100	0.0145%
Novel miRNA	81	107.7	8,727		0.000005%

4.5.3.3 Identification of other small ncRNAs

Except tRNAs/rRNAs, some other types of small ncRNAs also have some specific function for each species, such as snRNAs playing roles of processing of pre-mRNA, miRNAs and snoRNAs. Therefore, we also try to annotate them in the draft genome. The following are the methods we used to identify these small ncRNAs.

The snRNA genes are predicted using INFERNAL [148] against the Rfam database (release 11.0). In order to accelerate the speed, we performed a rough filtering prior to INFERNAL; by BLASTN against the Rfam database under E-value 0.01.

For the prediction of known miRNAs, we first aligned the mature miRNA sequences of *Arabidopsis thaliana/lyrata*, *Brachypodium distachyon*, *Medicago truncatul*a, *Nicotiana tabacum*, *Oryza sativa*, *Sorghum bicolor*, *Vitis vinifera* and *Zea mays* from the miRBase [149] (release 19) using MiRcheck [150] against the draft genome, allowing only one mismatch. The potential to form secondary structures by these miRNA candidates with their flanking region is checked later by RNAfold [151]. In order to identify novel miRNA for oil palm, RNA samples from flower (female and male), pollen, root, kernel (2.5 month) and mesocarp (1.5,2.5,3.5,4.5,5.5 months) are also collected for smRNA sequencing. First, rRNA-, tRNA- and known miRNA-related reads are removed. Then, regions having match reads, and flanking regions in the draft genome, are selected for novel miRNA prediction. If any of these regions and flanking regions can form a potential secondary structure, it is considered to be a potential novel miRNA. The remaining loci with small-RNA read hits, are used to predict novel miRNAs. Similar to known miRNA prediction, their flanking regions in these loci are also checked to see whether they can form potential hairpin loop structures.

The C/D snoRNAs are predicted using snoScan [145] with the yeast rRNA 18 and 25 methylation site and yeast rRNA sequences provided by the snoScan distribution. The minimum cutoff score is based on the default settings with 30 bits. Similarly, H/ACA snoRNAs are detected by snoGPS [152] using the yeast score tables and target pseudouridines; c.f. Table 4.11.

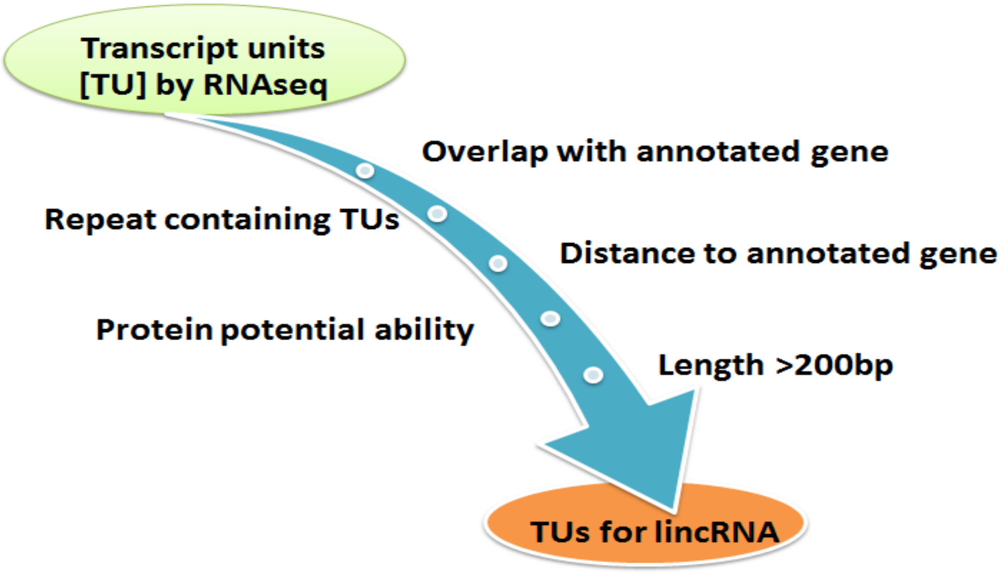
Table 4.11 lists all the predicted non-coding RNA genes in the draft oil palm genome. Among them, 199 known miRNA families have been identified; around 50% of them have been verified by our smRNA sequencing data for oil palm.

4.5.3.4 Identification of long intergenic noncoding RNA (lincRNA)

Long noncoding RNAs (lncRNA) are transcribed RNA molecules greater than 200 nucleotides in length. Based on their location in the genome, they are further divided into: (i) long intergenic noncoding RNAs (lincRNAs); (ii) long intronic noncoding RNAs (incRNAs); and (iii) natural antisense transcripts (NATs). Genomes of human, mouse and fly have been shown to encode lncRNAs that play important roles in cell differentiation, immune response, imprinting, tumor genesis and other important biological processes [85, 153]; besides, genetic mutations of human lncRNAs have been shown to be associated with diseases and pathophysiological conditions [154]. For plants, genome-wide search for ncRNAs has been previously conducted in *Arabidopsis thaliana* [5], *Medicago truncatula* [155], *Zea mays* [156] and *Tritucum aestivum* [157]. These lncRNAs show tissue-specific expression, and a large number of them are responsive to abiotic stresses. However, the function of these lncRNAs remains largely unexplored. Genomic loci of many lncRNAs are associated with histone modifications and DNA methylations suggesting an epigenetic regulation of these loci [158].

Hence, considering their important functions in the whole genome, we also attempt to identify and annotate lncRNAs in our Dura draft genome. The pipeline is similar to that of our previous work on *Arabidopsis thaliana* [5]. Due to the non-strand-specific sequencing for our datasets, only long intergenic RNAs (lincRNAs) are considered in our study here.

Figure 4.7 Pipeline for identification of long intergenic noncoding RNA


Transcript units [TU] by RNAseq
Overlap with annotated gene
Repeat containing TUs
Distance to annotated gene
Protein potential ability
Length >200bp
TUs for lincRNA

In general, all intergenic transcripts which can be transcribed can be considered as potential lincRNAs; however, some of them may also be related to other types of transcripts, such as truncated mRNAs, by-products of protein-coding genes, expressed repeats, or others. Such transcripts may confound the analysis of bona fide lincRNAs. Therefore, to facilitate further investigation of lincRNAs, we use the following criteria to provide a strict definition for lincRNAs for our oil palm sample:

(1) The transcript length must be at least 200 nucleotides;

(2) The transcript must contain no open reading frame (ORF) encoding >100 amino acids;

(3) The transcripts encoding lincRNAs must be located at least 500 bp away from any known protein-coding genes and genes for housekeeping ncRNAs;

(4) The transcripts must not encode any transposable elements (TEs);

Based on these features, to identify lincRNAs in our Dura genome, we subjected 29 RNA libraries derived from mesocarp, endosperm, root, leaf, flower and pollen to RNA-seq. Each RNA library yields around 20 million 101bp pair-end sequences; c.f. Figure 4.7. The total number of sequencing reads approaching 1 billion is comparable to or even higher than those reported by several RNA-seq studies in other species. The detail process for the identification can be summarized as:

First, these RNA sequence reads are aligned to our draft genome using Tophat [159, 160] and SAMtools [161]. The mapped sequences are then assembled into transcripts using Cufflinks [160] and Cuffcompare, yielding 13,204 to 27,434 transcripts in each organ. Of these, 10,524 to 21,715 transcripts (80%) are mapped to the genomic regions of annotated oil palm transcripts; c.f. Table 4.9. The remaining transcripts, using Cuffcompare, derived from intergenic regions are merged into 4,181 transcripts, which are potential lincRNA candidates.

Furthermore, by restricting the distance to annotated genes and length of the transcripts, around 3,000 lincRNAs are identified at the end.

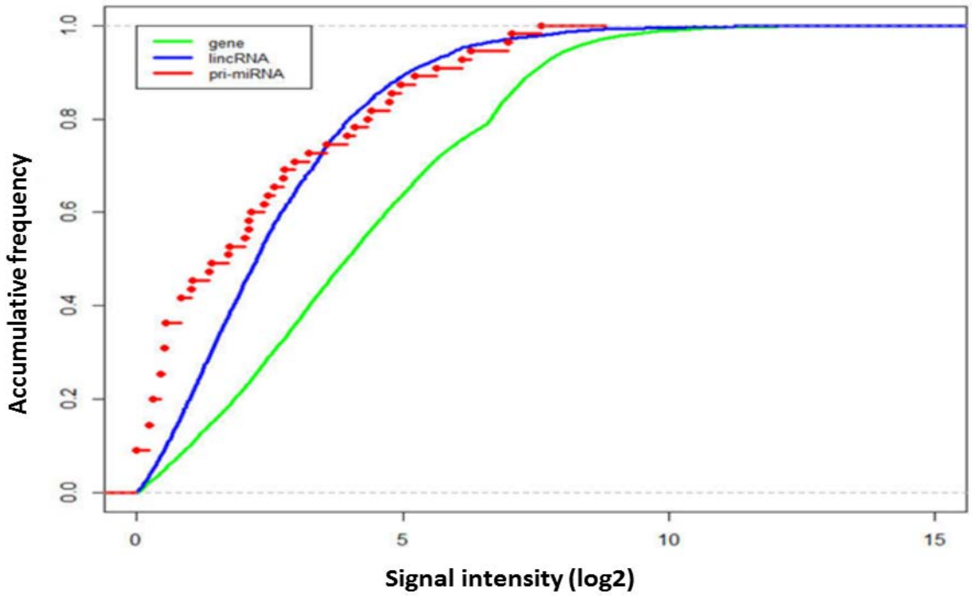
The whole pipeline for this process is summarized in Figure 4.7. The detail number of lincRNA on each tissue is presented in Table 4.12, which shows that pollen tissue may be a tissue different from others, because it has far fewer lincRNAs.

Table 4.12 Statistic information for the gene, lincRNA and miRNA identified by RNA seq data set

tissue	gene	gene_RNAseq	gene_repeat	miRNA	lincRNA	lincRNA_repeat
KD1.5	7,555	1,890	14,139	21	1,353	1,847
KD2.5	7,407	1,808	14,237	28	1,523	1,935
KD3.5	6,120	1,434	12,526	24	1,301	1,693
MD1.5	7,163	1,912	13,833	26	1,375	1,835
MD2.5	7,165	1,845	13,806	28	1,381	1,851
MD3.5	6,814	1,803	13,347	23	1,305	1,816
MD4.5	6,584	1,839	13,527	34	2,367	2,212
MD5.5	6,862	1,898	13,595	40	2,622	2,417
Leaf	7,164	1,899	13,568	27	1,436	1,916
MF1	5,991	1,527	12,176	22	1,352	1,785
MF2	6,962	1,832	13,285	24	1,629	2,008
FF1	6,136	1,505	12,373	25	1,190	1,695
FF2	6,075	1,451	11,959	23	1,061	1,618
Pollen	3,416	830	7093	15	822	1,028

Next, by comparing the expression level between protein coding genes, pri-miRNAs and lincRNAs in Figure 4.8, we see that the expression level of lincRNAs is lower than protein coding genes, but a little higher than pri-miRNAs. These results suggest that lincRNAs may differ from mRNAs in their biogenesis, processing, and/or stability. Moreover, the relatively low expression level suggests that very few of them are detected by cDNA/EST library.

Figure 4.8 Expression level of protein coding gene, pre-miRNA and lincRNA



4.6 Gene family for fatty acid pathway

Oil palm is a highly efficient oil-producing crop. The detail mechanism is still unknown. Thus, after constructing the Dura draft genome, it is interesting to investigate genes involved in lipid biosynthesis pathways (Table 4.13) and do comparative studies between different species (*Arabidopsis thaliana*, date palm, *Vitis vinifera*, *Glycine max* and *Oryza sativa).*

We have summarized all the genes related to lipid biosynthesis pathways from different species in Table 4.13. These results show that oil palm and soybean have the highest copy number for lipid-pathway genes, which may explain why these two species accumulate the highest amount of fruit/seed oil. In addition, oil palm has more *FAD* genes than soybean, which are responsible for transferring oleic acid to palmitic acid. Although the detail relationship between these family members is still unknown, we believe that the huge number of lipid-related genes in oil palm and soybean can play different roles related to lipid synthesis under various conditions.

Table 4.13 The number of genes in fatty acid biosynthesis pathways for each plants

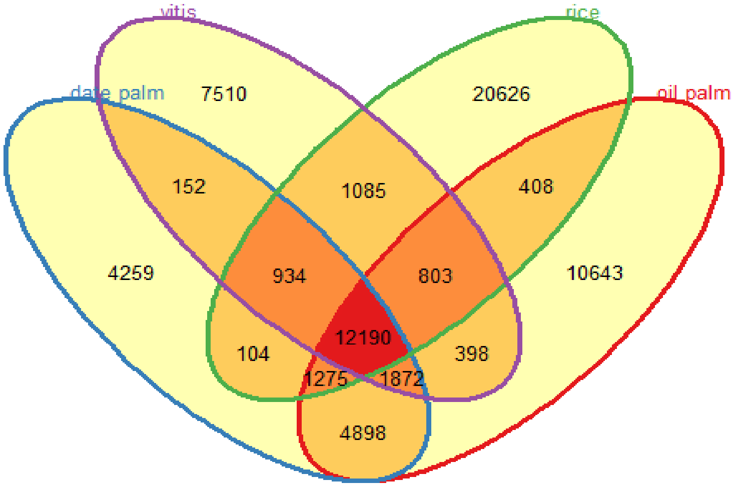
	Arabidopsis	**oil palm**	date palm	Vitis	**soybean**	rice
ACC	1	6	3	3	4	0
DGAT	1	4	3	2	5	2
EAR	0	7	3	0	8	4
FAD	2	**11**	1	0	**0**	0
FAT	0	11	4	5	12	6
GPAT	10	7	0	6	28	24
HAD	2	3	0	1	4	6
KAR	1	5	2	4	4	10
KAS	27	49	21	27	51	32
LACS	0	3	3	5	13	5
LPAAT	0	4	1	0	0	0
MAT	1	2	2	1	2	1
PAP	0	3	2	0	0	0
PDAT	1	6	0	2	4	1
PDH	4	17	17	6	14	5
PK	0	31	19	14	30	10
SAD	7	8	5	10	5	9
total	57	**177**	86	86	**184**	115

4.7 Homologous genes

Comparative homolog analysis, including date palm [119], *Vitis vinifera* [35] and *Oryza sativa* [118], suggests that there are around 36,015 protein-coding genes in oil palm, with 12,190 protein-coding genes being shared by date palm, *Vitis* and rice; see Figure 4.9. Among them, date palm shares the most number of homologs (4,898) with oil palm, much more than the 408 with rice, 398 with grape (*Vitis*). This is consistent with their close evolutionary relationship.

In addition, there are 10,463 unique protein-coding genes in oil palm; some are potentially employed in important biological processes (for example, the control of flowering time or secondary metabolisms).

Figure 4.9 Venn graph of homologs between oil palm, date palm, Vitis and rice



4.8 Whole-genome duplication

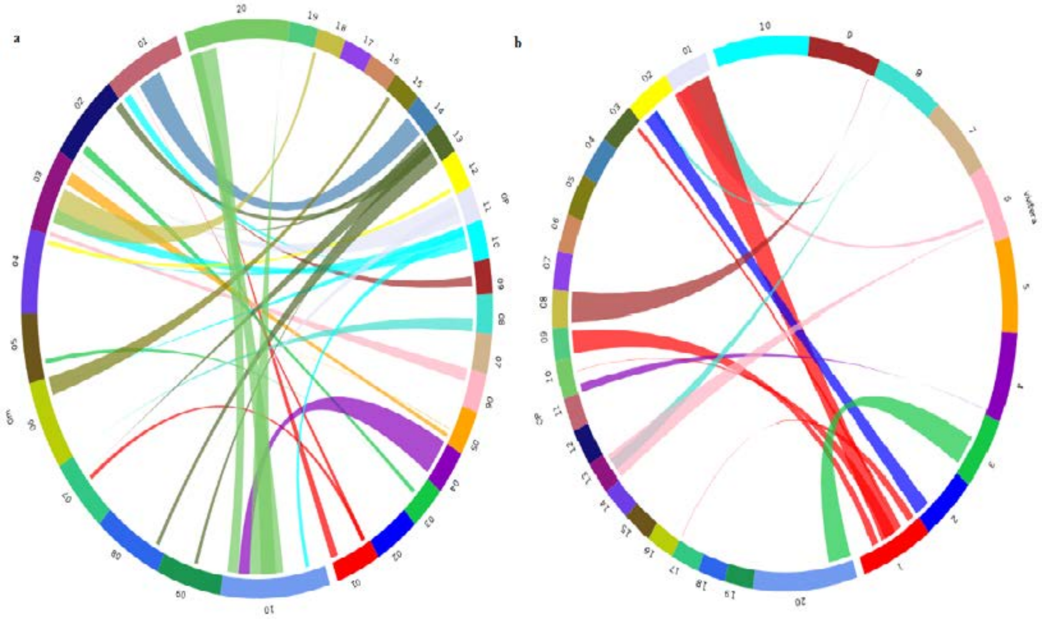
Genome-wide duplication in angiosperms is common, and represents an important molecular mechanism that has shaped modern plant karyotypes.

To generate a pair-wise alignment of gene models between oil palm and *Vitis*, oil palm and soybean, oil palm and date palm, all predicted genes are aligned to the reference genes by Mummer [101]. The criterion is that the number of genes in one synteny block should be more than 5. In order to clearly visualize these synteny blocks, we just selected the 10 longest chromosomes from *Vitis* and rice, and the 20 longest scaffolds from date palm and oil palm.

Each homolog is shown as a black dot, while a synteny region is represented in a rectangle in Figure 4.12. From this figure, we can see the conservation regions between oil palm and the species mentioned, which shows that soybean shares more conservation regions than the other two plant species. Similar results can be found in Figure 4.10.

One possible reason for this phenomenon is that the annotation of the soybean genome is much better and the top 10 chromosomes of the soybean genome are much longer than others, especially date palm.

Figure 4.10 a: synteny region between oil palm and soybean b: synteny region between oil palm and Vitis



Under a circle view for these synteny regions, it is clear that the synteny regions between *Vitis* and oil palm are located in chromosome 1 of *Vitis*, detail synteny locations for these two chromosomes are shown in Figure 4.11.

Figure 4.11 Detail synteny regions for one chromosome from oil palm

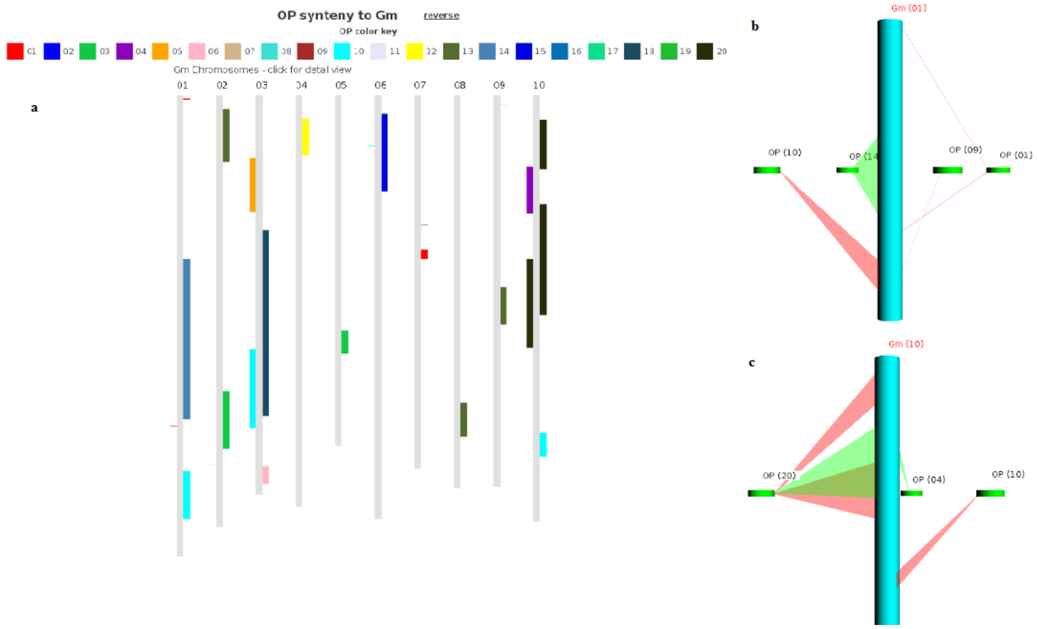
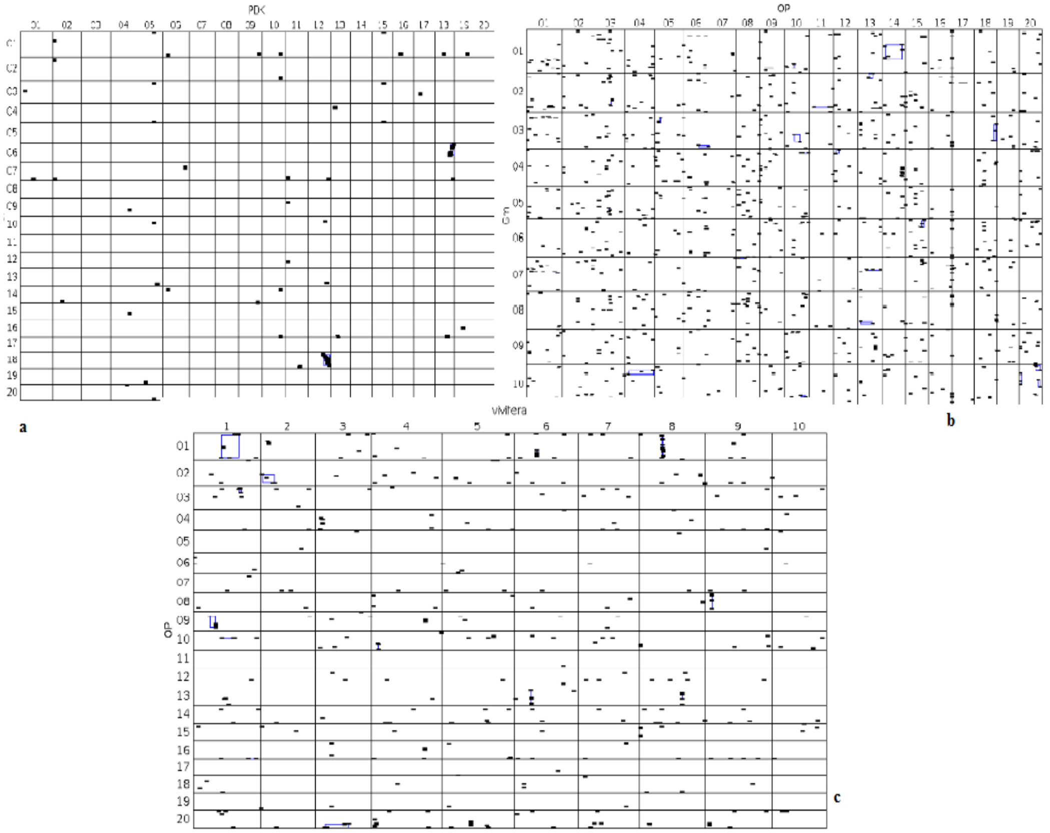


Figure 4.12 The synteny region in the detail location of each chromosome. a Synteny region between oil palm and date palm b Synteny region between soybean and oil palm c Synteny region between oil palm and *Vitis*



4.9 Evolution history of oil palm

Due to the difficulty of constructing transgenic oil palm, the rapid growth of oil yield has been stimulated in major part by progress in research and development (R&D). Discovery of the single-gene inheritance for shell thickness and subsequent adoption of D X P planting materials has led to a quantum leap in oil-to-bunch ratio from 16% (Dura) to 26% (Tenera). Thus oil palm cultivation becomes more profitable.

Further yield improvements have subsequently been made through breeding for Dura and Pisifera with specific combining ability. During the past thirty years, modern breeding methods based on quantitative genetics theory have been extremely successful in improving oil productivity. Hence, in our study, based on our draft genome, we also attempt to identify the most common alleles at the majority of polymorphic sites in the genome and provide some evidence and suggestion for future breeding.

Benefitting from next-generation sequencing, a wide range of genetic and archaeological studies have been carried out to examine the phylogenetic relationship with other species, like rice [162]. Molecular genetic analyses indicated that *indica* and *japonica* originated independently. Meanwhile, population genetics analyses of genome-wide data of cultivated and wild rice have also tended to suggest that *indica* and *japonica* genomes generally appear to be of independent origin [3]. Despite these advances in other species, there is still a lack of clarity of the evolutionary history of oil palm domestication by population-scale whole-genome sequencing. An in-depth investigation of the haplotype structure near the domestication sites is critical for evaluating the direction of introgression. Moreover, a comprehensive map of oil palm genome variations can facilitate genetic mapping of complex traits in oil palm.

Table 4.14 Description of 12 oil palm strains

Sample	Origin-species	coverage
TS1	AVROS-pisifera	2.86
TS2	EKONA-pisifira	2.82
TS3	GHANA-pisifera	3.14
Dura A	Asian-Dura	3.04
Com1 DeLi1	Com1 DeLi1-Tenera	2.57
Malaysia 08	Malaysia 08-Tenera	3.47
LT2O3	LT2O3-Tenera	3.37
T2BIS2	T2BIS2-Tenera	2.83
Com NiG 02	Com NiG 02-Tenera	3.44
Com Gha 04	Com Gha 04-Tenera	3.04
Dura B	Dura B (Asia)-Dura	3.09
AGO T 08g	AGO T 08g-Tenera	3.69

Hence, after constructing the draft genome for Dura, we also collect diverse oil palm strains from the whole world for sequencing and attempt to carry out genome-wide association studies for many agronomic traits in oil palm evolutionary history.

From the large collection of oil palm in the world, we select 3 categories including 12 different strains mainly from Asia and African, spanning the native geographic range of the species; c.f. Table 4.14. From the coverage of each sample, it can be found that the sequencing depths for whole genome is around 2-3 fold.

4.9.1 Overview of diversity for oil palm

In order to find SNPs between these oil palm strains with our draft genome, the pair end reads of all the samples are first aligned against our draft Dura genome. After alignment, SNPs between our reference genome and other oil palm strains are called by SAMtools [161], c.f. Table 4.15. By comparing the number of SNPs with the reference genome, it is found that DuraA, DuraB and Malaysia shares little SNPs with other species, which is consistent with the fact that all of them are from Asian countries.

Table 4.15 SNP number between each oil palm strains and reference genome

strains	total SNPs
TS1	3,435,595
TS2	3,316,953
TS3	3,801,044
Malaysia	2,645,361
DuraA	383,200
DuraB	2,379,826
LT2O3	3,840,078
T2BIS2	2,596,119
Gha	3,286,761
DeLi	2,061,389
AGO	4,196,381
NiG	3,883,414

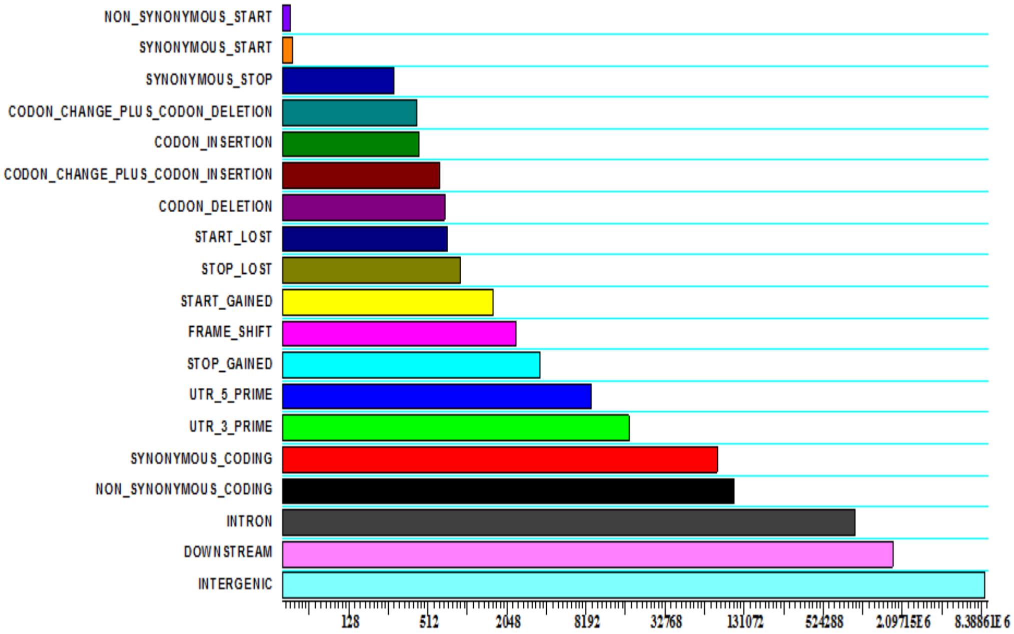
In order to explore various information of SNP for oil palm, we compare the following information for different groups [Figure 4.13].

a. location information for each SNP: Intergenic region, UTR, intron, exon, CDS, downstream (length: 5Kb), upstream (length 5Kb).

b. coding feature: NON_Synonymous_coding (SNP causes a codon that produces a different amino acids), Synonymous_coding (SNP causes a codon that produces the same amino acid).

c. codon level: Codon_change (one or many codons are changed), Codon_Insert (One or many codons are inserted), Codon_deletion (one or many codons are deleted), Exon_deleted (A deletion removes the whole exon), Start_Lost (start codon is mutated into a non-start codon), Synonymous_start (start codon is mutated into another start codon), Synonymous_stop (stop codon is mutated into another stop codon), Stop_lost (stop codon is mutated into a non-stop codon).

Figure 4.13 Statistic for different SNP categories of oil palm

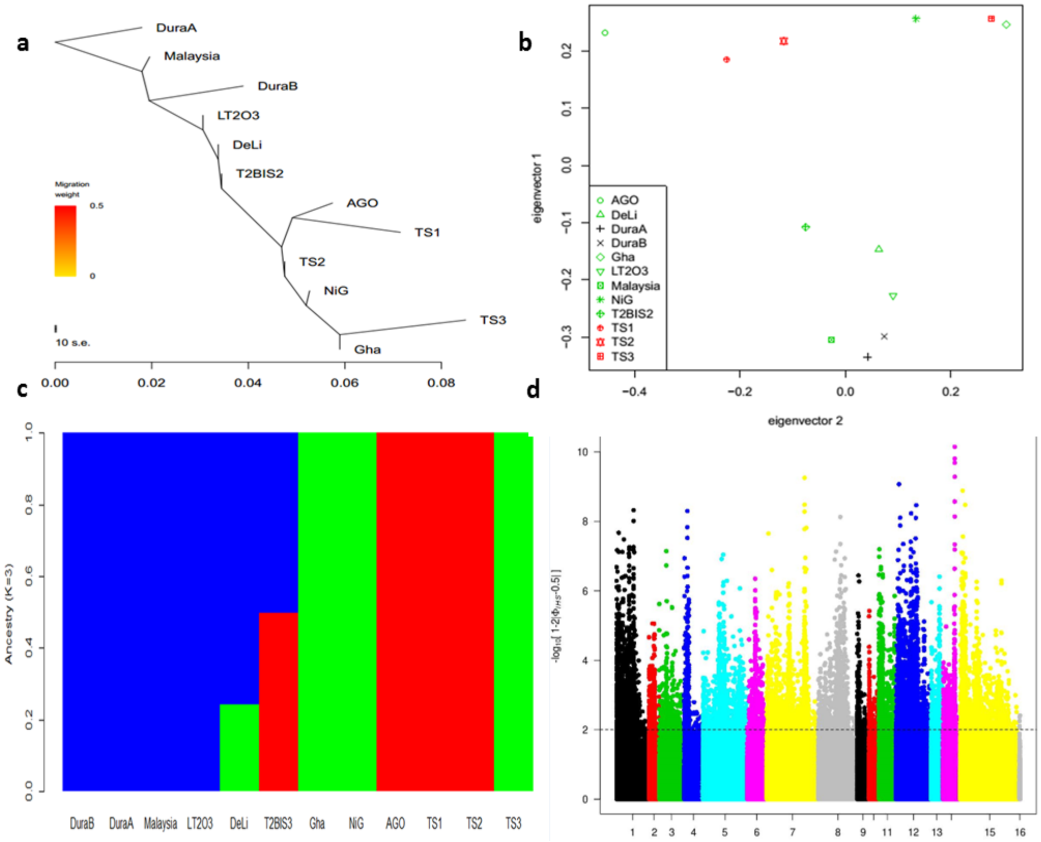


From the distribution for the SNP numbers in all categories [Figure 4.13], we find that most (72%) are located in intergenic regions, and only a few of them are located in coding sequence regions. Among the latter, there are 110,446 nonsynonymous SNPs and 81,774 synonymous SNPs. Thus the ratio of nonsynonymous-to-synonymous substitutions is 1.35, which is similar to rice genome (1.29) [3], higher than *Arabidopsis* (0.83) [163] and lower than soybean (1.61) [4]. In addition, we have also identified more than 1,000 stop or start codon change-related protein coding genes. As for the biological reason behind these start/stop codon mutations, it needs more detail exploration of these genes.

4.9.2 Structure and population analysis for oil palm

Based on SNP data, we next investigated the population structure of these oil palm strains to understand their evolutionary relationship. On the basis of the neighbor-joining tree, Figure 4.14.a, same with our observation by SNP number, DuraA/B and Malaysia are the nearest neighbors to our reference genome, which are totally different from the Pisifera group. For the Tenera group, it displays some divergent phenomenon, different from Pisifera/Dura group. This evolution tree provides us a good clue to select crossing species for breeding. The further from the reference strain, the more chance of getting a good crossing outcome, because there is more possibility to get a genome recombination.

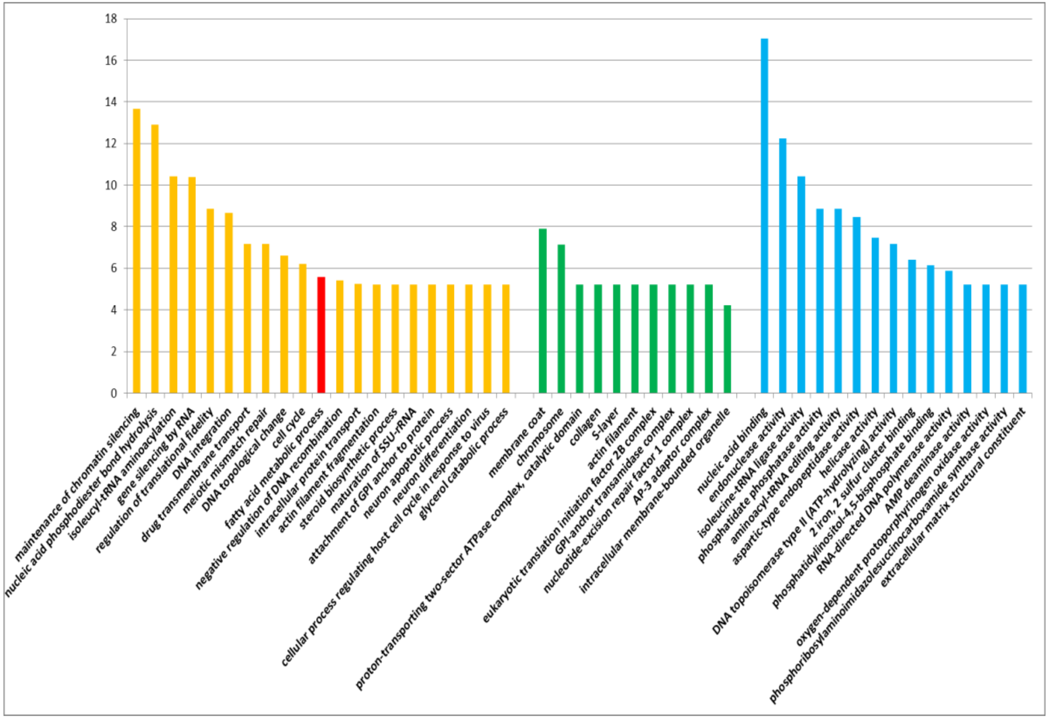
Figure 4.14 Population genetic analysis of oil palm a: neighbor-joining tree for 12 different oil palm strains b: PCA result for 12 different oil palm strains c: Bayesian clustering (STRUCTURE, K=3) d:iHS score for different diversity sites across all chromosomes



Similar results can also be obtained from principal component analysis (PCA) analysis, which shows that DuraA, DuraB and Malaysia are the nearest; c.f. Figure 4.14.b. Interestingly, the Tenera group is classified into several groups. That may be because Tenera are crosses between Dura and Pisifera group, some of them may be similar with maternal line, whereas some of them may be similar with the paternal line, and others may have their own features. These results are also supported by the Bayesian clustering program STRUCTURE [164], with K=3 [Figure 4.14.c].

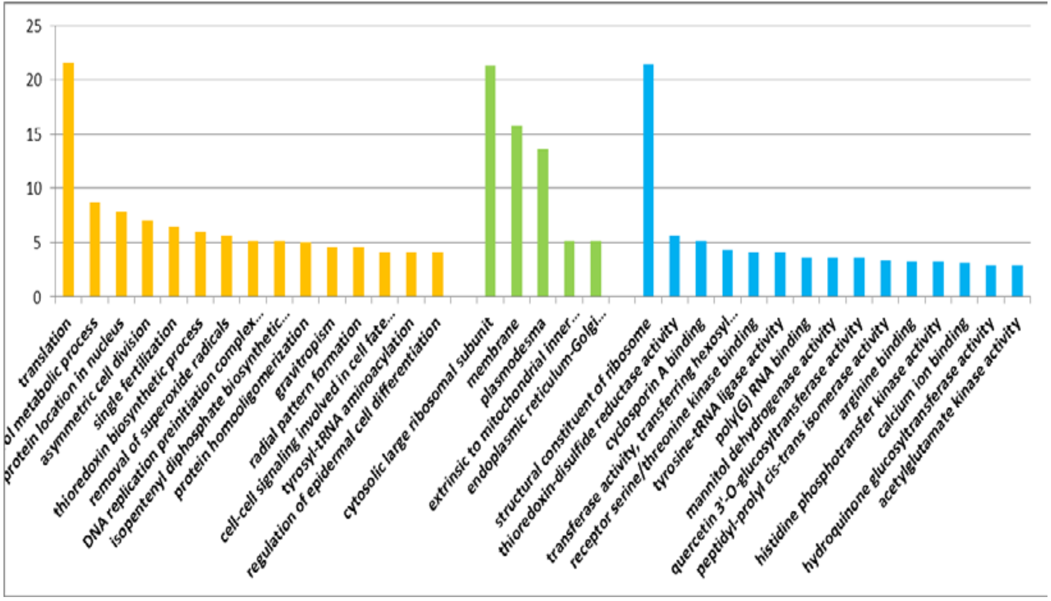
Nucleotide diversity is a common measure of genetic variation. It is usually associated with other statistical measures of population diversity, and is similar to expected heterozygosity. This statistic may be used to monitor diversity within or between ecological populations, to examine the genetic variation in crops and related species or to determine evolutionary relationships. The integrated haplotype score (iHS) is a measure of the amount of extended haplotype homozygosity (EHH) at a given SNP along the ancestral allele relative to the derived allele. This measure was designed by Voight et al as a method to describe a recent map of positive selection in the human genome [165]. In our study, the iHS score across the whole genome are also explored [Figure 4.14.d].

Figure 4.15 Enriched GO terms for high-diversity gene locus Orange: biological process Green: cellular component Blue: Molecular function



After selecting high- and low-diversity locus using iHS score, by GO term enrichment for these high- and low-diversity genes, we find that gene families with essential functions (for example, translation, maintenance of protein location in nucleus) tend to have substantially lower substitution ratios [Figure 4.16], whereas gene families that function in regulatory processes, such as fatty acid metabolic process and steroid biosynthetic process, have higher ratios [Figure 4.15].

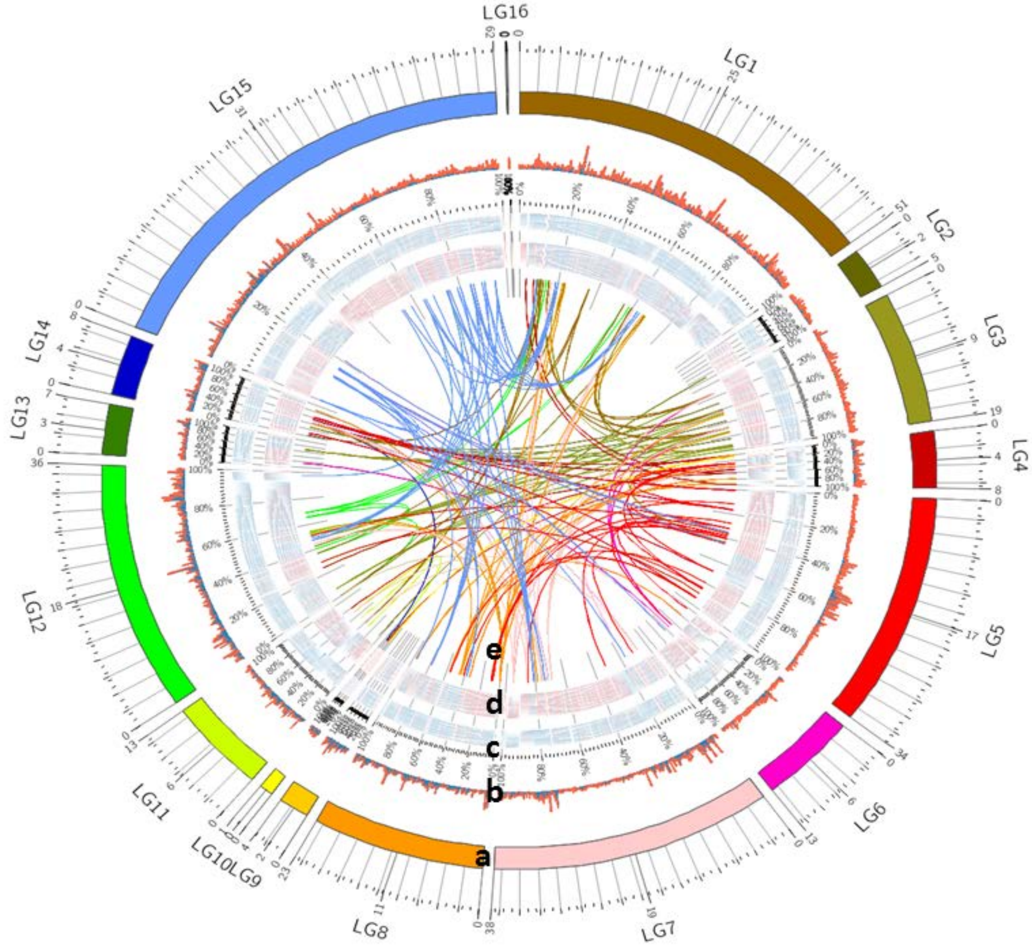
Figure 4.16 Enriched GO terms for low-diversity gene locus Orange: biological process Green: cellular component Blue: Molecular function



In summary, we provide new insights into how oil palm strains evolved by SNP analysis. We will further investigate the relationship between these SNP-based markers and genotypes of oil palm, which can guide future breeding efforts.

Totally, we have summarized all the basic genome information for oil palm in Figure 4.17.

Figure 4.17 Global overview about chromosome of oil palm   a: chromosome information b: iHS score distribution c: gene density d: repeat density e: segmental duplication in genome



4.10 Conclusion

In this chapter, we have applied our proposed reference-based genome assembly pipeline to genome sequencing data of Dura oil palm. From the results on our Dura sample, it is clear that our pipeline outperforms de novo assembly methods and other reference-based methods (ABACAS).

Evaluation from three independent methods---EST coverage, genome completeness and linkage map---has demonstrated the accuracy and completeness of our draft Dura genome. This is the first complete genome sequence for Dura, and is the second complete genome sequence for oil palm. Our draft genome can be used for downstream analysis.

Based on this draft genome, gene annotation, ncRNA annotation and lincRNA annotation are performed. This draft genome encodes around 30,000 protein-coding genes, 200 miRNAs and 1,000 lincRNAs. These annotations should facilitate research on oil palm.

By the statistics information of lipid-related genes and comparison with other oil crops, we also get a general overview of possible reasons for the high oil yield of oil palm.

By resequencing of 12 different oil palm strains, we have obtained a clearer overview of the evolutionary history of the oil palm family. The result provides some evidence and suggestion for improvement of oil palm by cross-breeding.

In summary, we believe our results provide a rich genome resource for molecular research and breeding of oil palm.

Chapter 5

VISUALIZATION OF VARIOUS GENOME INFORMATION

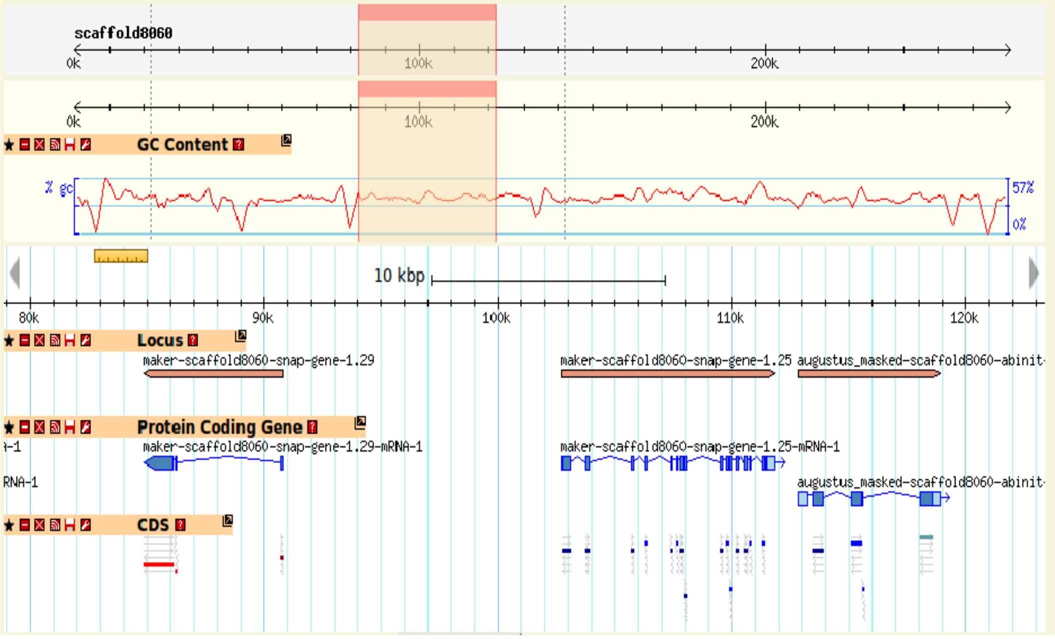
To provide convenient access and query for the research community, especially biologists, we have built several visualization tools for various genome information. Based on the characteristics of various genome information, our database aims to provide the following 5 essential functions: (1) Visualization of location and structure information for each transcription units, such as protein-coding gene, miRNA and long noncoding RNA; (2) Expression levels from various source, such as RNA-seq, tilling array and Chip-seq; (3) Epigenetic modifications information (e.g. DNA methylations and histone modifications) across genomic regions; (4) A collection of siRNA sequencing dataset across the whole genome; and (5) BLAST function to support homolog search.

Therefore, we integrated all the genome information of Dura oil palm into our GBrowse-based platform that was used for another long noncoding RNA database [PLncDB][137]. Here, we explain these functions in detail in following sections.

5.1 An online database to deposit, browse and download genome element

We constructed a database [Figure 5.1] using the open source GBrowse library [166] to integrate and visualize different sources, such as protein coding gene, small ncRNA and lincRNA annotations. In addition, a list of de novo gene prediction results from Augustus [68], SNAP [133] and Tophat [160] and final gene model integrated from Maker 2 [138] can also be visualized [Figure 4.5]. As for expression information, we adopted a new file format BigWig [167] similar to BAM, which is binary, compressed and reduces loading time to the browse. The database can be accessed or queried in various ways. Specific searches can be performed using the name/keywords of gene/protein and/or location on the chromosome. At the same time the entire database is available for download in different formats.

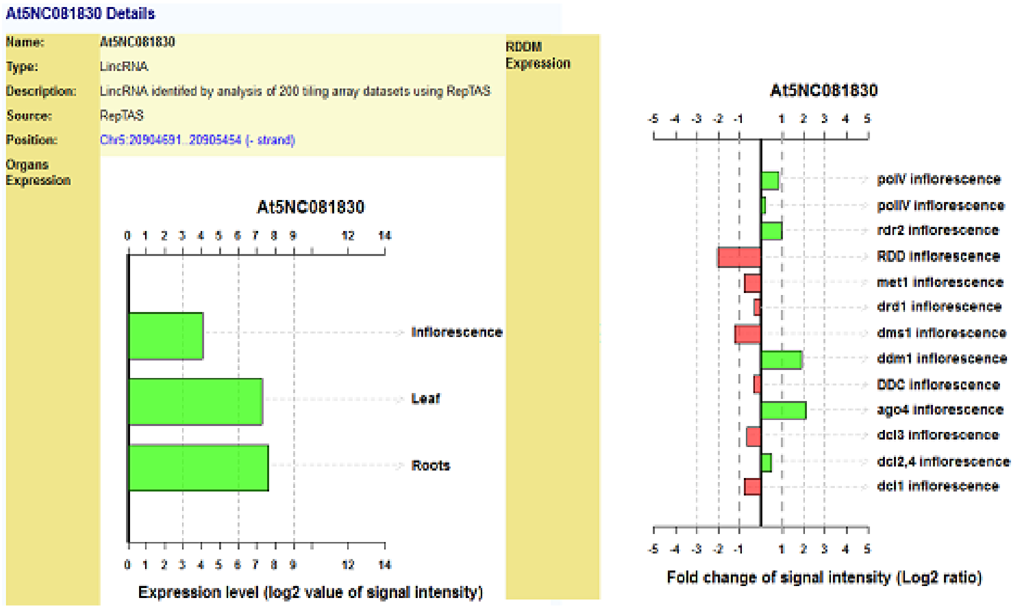
Figure 5.1 Snapshot of the GBrowse database to visualize the genome element


scaffold8060
0k
100k
200k
GC Content
% gc
57%
0%
10 kbp
80k
90k
100k
110k
120k
Locus
maker-scaffold8060-snap-gene-1.29
maker-scaffold8060-snap-gene-1.25
augustus_masked-scaffold8060-abinit
Protein Coding Gene
maker-scaffold8060-snap-gene-1.29-mRNA-1
maker-scaffold8060-snap-gene-1.25-mRNA-1
augustus_masked-scaffold8060-abinit
RNA-1
CDS

5.2 Visualizing detail information for transcript unit

Just by clicking on a specific item, researchers can visualize mutant/stress related information; see Figure 5.2. By viewing these detail information, researchers can obtain potential function of genes, according to the relative expression level in each tissue. By experimental verification, biologist can conduct further mechanisms-related research, potentially saving a lot of time instead of screening for different gene candidates.

Figure 5.2 An example of detail information for transcript unit in the database



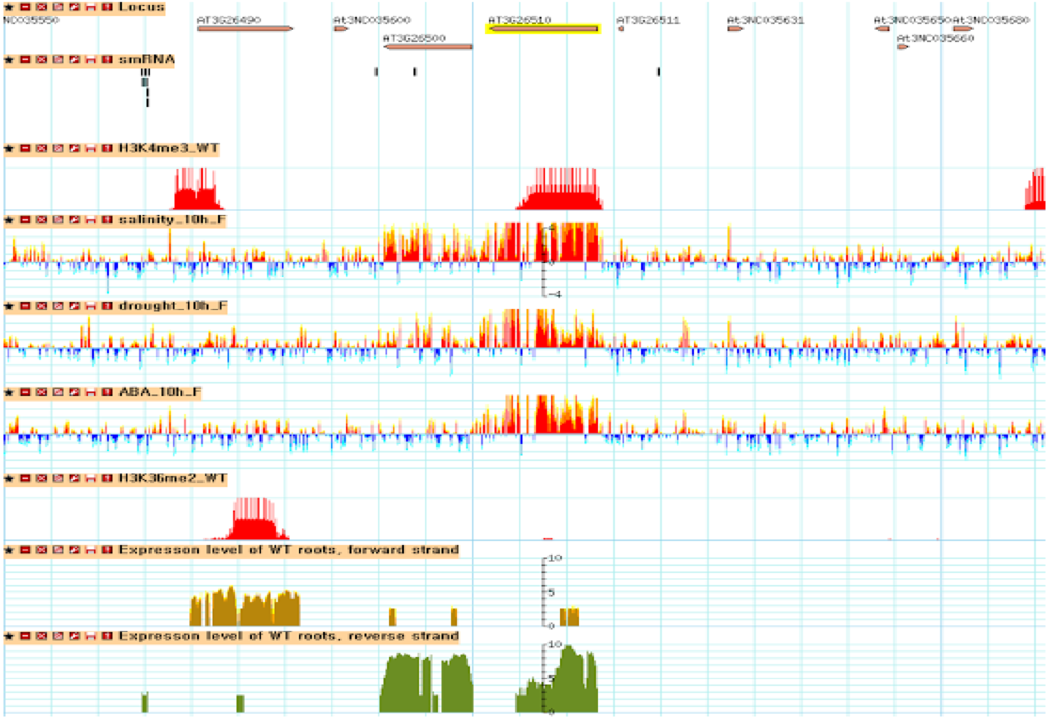
At5NC081830 Details

Name:	At5NC081830	RDDM Expression
Type:	LincRNA	
Description:	LincRNA identified by analysis of 200 tiling array datasets using RepTAS	
Source:	RepTAS	
Position:	Chr5:20904691..20905454 (- strand)	
Organs Expression		

5.3 Visualizing relative expression level across the whole genome

Besides the location information, we also provide the relative gene expression level for the whole genome. The expression level is measured by RNA sequencing technique, Chip-seq and array platform, which divides the whole genome into equal small window sizes [Figure 5.3]. By this method, the user can have a clear view about the activity and epigenetic information for the whole genome, even the intron/exon difference. Like the example in Figure 5.3, we can see that this protein coding gene is highly up-regulated under salt, drought and ABA treatment, which means this gene may have specific functions for abiotic stresses.

Figure 5.3 Snapshot for the expression level of our database



5.4 Visualizing smRNA abundance across the whole genome

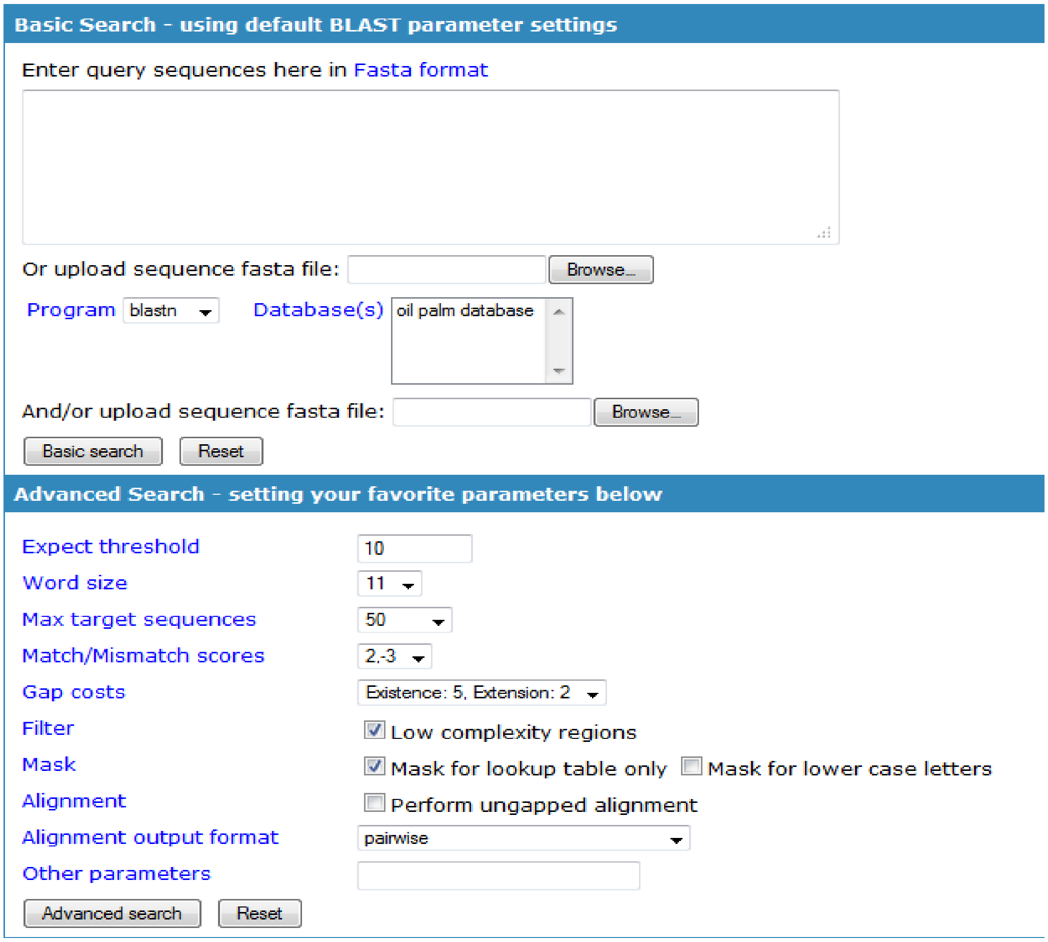
Noncoding RNAs such as ribosomal RNAs, transfer RNAs, small nuclear RNAs, small nuclear RNAs, and small interfering RNAs, can serve catalytic and scaffolding functions in transcription, messenger RNA processing, translation, and RNA degradation [152]. Besides the location for various ncNRA families, we also provide the expression level for all these ncRNAs based on our smRNA sequencing datasets [Figure 5.3]. In this case, miRNAs related to different conditions/mutants can be easily found and queried. Moreover, combined with epigenetic information in the whole genome, it is even possible to find some siRNA-medidated epigenetic silence locus.

5.5 BLAST tool

Another requirement and use of our genome resource is homology search by sequences from other species. Therefore, in addition to GBrowse-based tool for visualizing oil palm genome information, we also support BLAST function for any given query sequence [Figure 5.4].

For this BLAST tool, we have enabled querying using any nucleotide and peptide sequences. At the same time, user also can conduct nucleotide-level homolog search and protein-level homolog search by selecting BLASTN, BLASTP, BLASTX, etc.

Figure 5.4 Snapshot of the BLAST function for oil palm database



Basic Search - using default BLAST parameter settings

Enter query sequences here in Fasta format

Or upload sequence fasta file:

Program: blastn

Database(s): oil palm database

And/or upload sequence fasta file:

Advanced Search - setting your favorite parameters below

Expect threshold	10
Word size	11
Max target sequences	50
Match/Mismatch scores	2,-3
Gap costs	Existence: 5, Extension: 2
Filter	☑ Low complexity regions
Mask	☑ Mask for lookup table only ☐ Mask for lower case letters
Alignment	☐ Perform ungapped alignment
Alignment output format	pairwise
Other parameters	

5.6 Conclusions

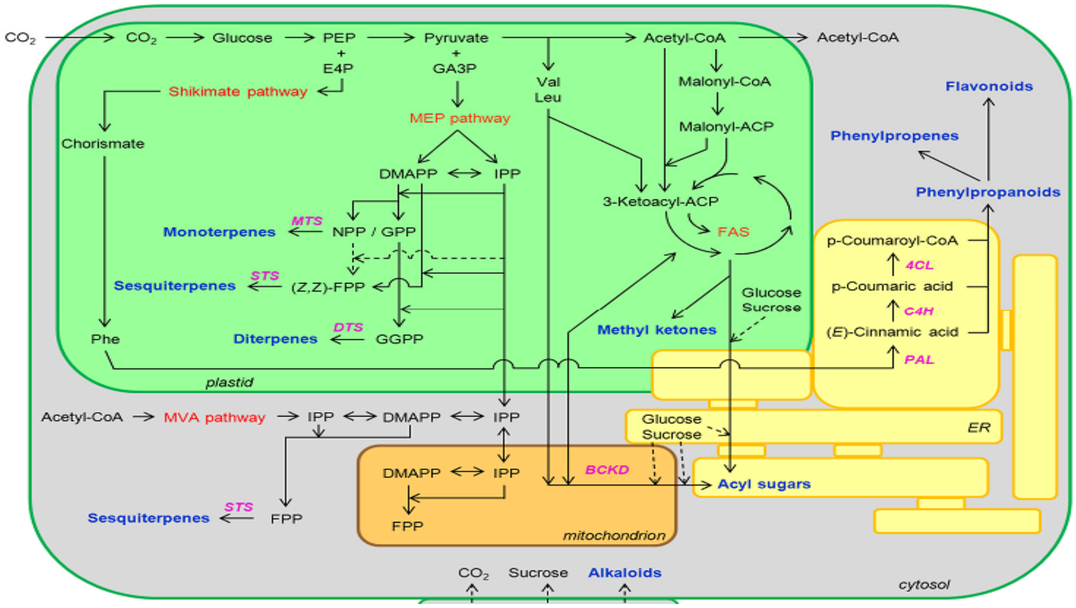
In this chapter, two useful tools---a GBrowse-based database and a BLAST tool---have been developed. Using these two tools, people, especially biologists can easily guess the potential function for specific genes and design experiments to verify their hypothesis. This should accelerate oil palm research.

Chapter 6

WEIGHTED PATHWAY APPROACH

Different from primary metabolites, secondary metabolites are another important group for plants. Although they do not play essential functions, like lipids as one of the sources of energy, they typically mediate interactions of plants with other organisms. These interactions include those of plant-pollinators, plant-pathogens and plant-herbivores. Although these interactions are not necessary for the basic life of plants, they are very useful between plants and the environment.

Figure 6.1 Simplified schematic overview of the biosynthesis of the main secondary metabolites stored and/or secreted by glandular trichome cells. Major pathway names are shown in red, key enzymes or enzyme complexes in purple, and stored and/or secreted compounds in blue. [168]

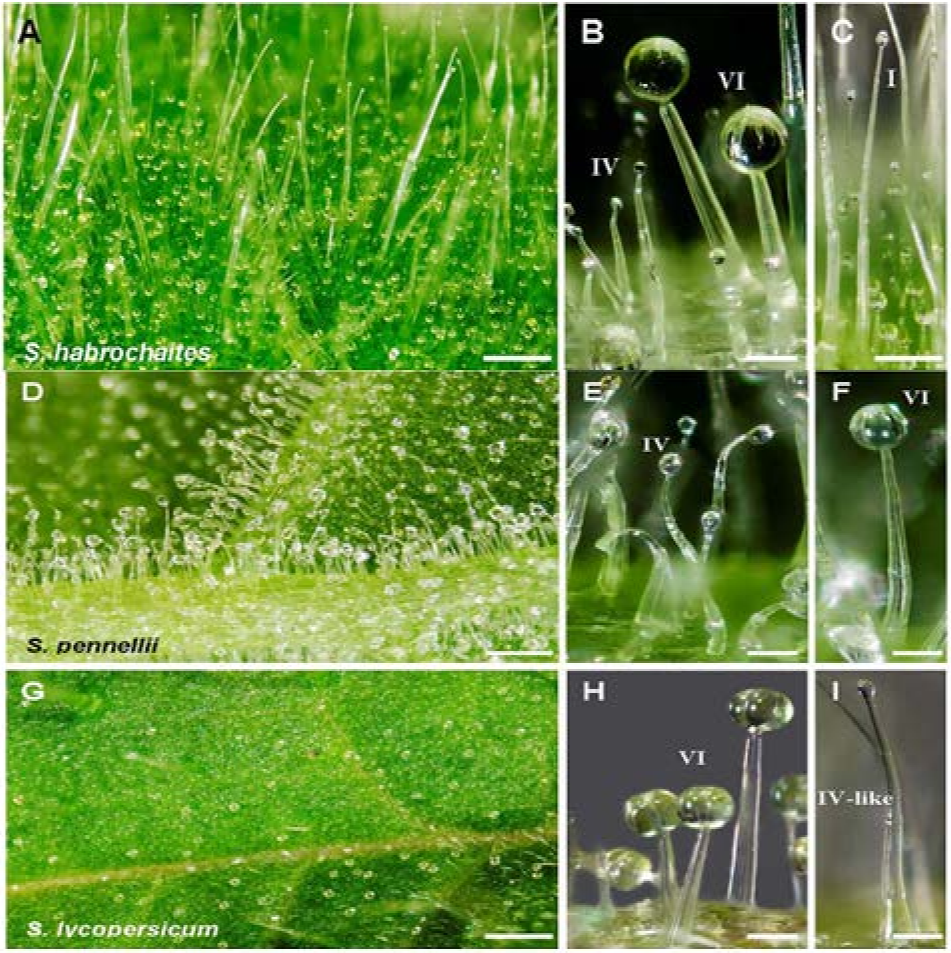


Commonly, secondary metabolites can be classified based on their chemical structures (for example, having rings, containing a sugar), compositions (containing nitrogen or not), their solubility in various solvents, or the pathways by which they are synthesized (e.g., phenylpropanoid, which produces tannins) [168]. A simple classification includes three main groups: Terpenes (made from mevalonic acid, composed almost entirely of carbon and hydrogen), phenolics (made from simple sugars, containing benzene rings, hydrogen, and oxygen), and nitrogen-containing compounds (extremely diverse, may also contain sulfur) [Figure 6.1].

Most of these secondary metabolites are produced by hair-like epidermal structures, commonly referred to as trichomes if they are present on the aerial parts [Figure 6.2]. Trichomes can be single-celled or multicellular, but the criterion that is most commonly used to classify them is whether they are glandular or not [168]. For the model plant *Arabidopsis*, only non-glandular trichomes can be found, which are unicellular and can be either unbranched, or has two or five branches [169]. These trichomes are polyploid and have been extensively studied with respect to their development [170]. However, large amounts of secondary metabolites are usually produced by glandular trichomes, which can be found on approximately 30% of all vascular plants.

Secondary chemicals of plants have important uses for humans. Most pharmaceuticals are based on plant chemical structures, and secondary metabolites are widely used for recreation and stimulation (the alkaloids nicotine and cocaine, the terpene cannabinol). The study of such plant use is called ethnopharmacology. Psychoactive plant chemicals are central to some religions, and flavors of secondary compounds shape our food preferences. The characteristic flavors and aroma of cabbage and relatives are caused by nitrogen-and sulfur-containing chemicals, glucosinolates, which protect these plants from many enemies [1]. The astringency of wine and chocolate are derived from tannins.

Figure 6.2 Glandular trichomes in section Lycopersicon. [168]



Despite the large commercial application of secondary metabolites, many of them are still harvested naturally. The accumulation of these specialized metabolites in plants is low and depends on environmental factors. Access to such compounds is often inadequate, and the reliance on the production of metabolites from naturally-growing plants is not always sustainable. Hence, in this chapter, we introduce a weighted pathway approach to investigate secondary metabolisms by next-generation sequencing techniques.

6.1 Background

To investigate the detail mechanism for these secondary metabolisms, the first problem is that of identifying or detecting secondary metabolisms and the associated protein-coding genes and metabolites. Only recently the monitoring of metabolites has grown into an ‘omics’ level field [171]. Gas chromatograph-mass spectrometry (GC-MS) has been applied to examine the effects of genetic and environmental manipulations [172], to determine phloem composition [173]. GC-MS is currently the most developed of the available analytical tools for metabolites. The growth of this technology offers an opportunity to view the effect of elicitation on metabolism at a larger scale than previously possible. However, GC-MS technique can only detect the relative expression level of different secondary metabolites. It is still unknown how to improve the yield of these secondary metabolites.

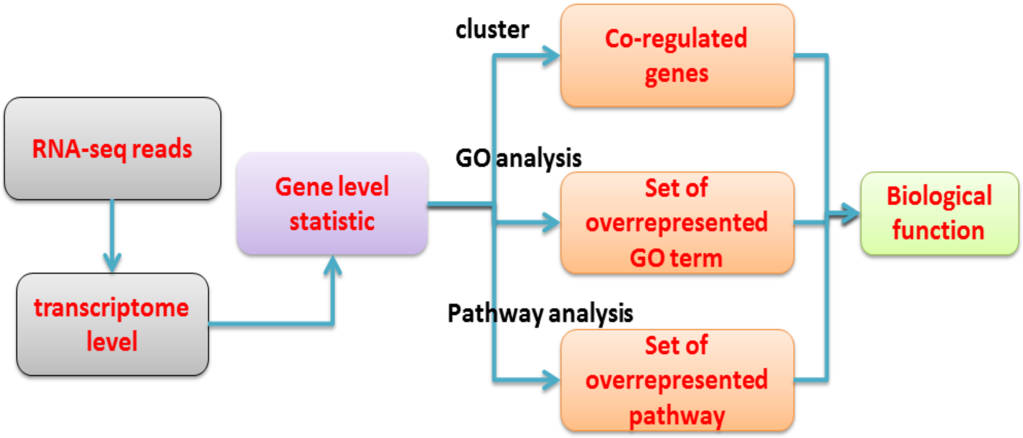
At the beginning, new metabolites are often discovered by homology-based cloning of genes involved in their biosynthesis [12, 13]. More than a decade ago, DNA microarrays have provided scientists the capacity to simultaneously investigate thousands of features in a single experiment. This capability has been exploited not only to monitor the steady-state expression of genes, but also to map the genome-wide binding sites of DNA interacting proteins (ChIP-on-chip) and to survey long-range DNA interactions (4C). The over-whelming wealth of knowledge generated by microarrays has created entirely new fields of research and, as the underlying technology became broadly adopted, microarrays forever changed the way in which high-throughput science is done.

However, because of the lack of extensive genomic data for the vast majority of plants, especially plants which are the major secondary metabolite producers, it is difficult to use the common microarray-based approach for transcriptome analysis. This is because such an approach requires prior designed probes for each target. The limitation of its prior requirement hinders its applications, especially for plants producing secondary metabolites.

Equally revolutionary technologies are currently emerging in the form of new methods of sequencing, termed massively parallel sequencing (MPS, also called next-generation/ultra-high-throughput sequencing). With the development of this technique, new genes/specific transcripts can be discovered and analyzed in a genome-wide model [14, 15], even without a reference genome.

Using transcriptome data produced by next-generation sequencing techniques, some interesting gene candidates can be identified by differential expression analysis between different conditions/tissues. However, for many investigators, the list of differentially expressed genes often fails to provide mechanistic insights into the underlying biology of the condition being studied [174]. In addition, people are also interested in new functions or compounds. Most previous studies, involving next-generation sequencing data, have just focused on the known secondary metabolisms to identify bottlenecks in known biosynthesis pathways [175-177]. Thus, in this way, the advent of high-throughput sequencing technologies presents a new challenge, that of predicting new functions or new metabolites for plant samples. For several years ago, there has been a paradigm shift from individual genes to gene sets [Figure 6.3]. Each of these gene sets performs a specific function. These methods can be classified and summarized into the categories below.

Figure 6.3 Analysis methods for RNA-seq data



6.1.1 Co-regulated genes

One approach to this challenge has been to construct co-expression networks [178, 179]. From these networks, some genes that are co-expressed with known proteins/metabolites in known biosynthesis pathways of some secondary metabolites are extracted. These genes are hypothesized to play important roles in the biosynthesis of those secondary metabolites [71]. However, most of these works just use some traditional statistical methods to identify co-expression pairs, and the resulting accuracy is very limited [180, 181]. In addition, most of these studies lack experimental results that verify or support the predictions and conclusions, especially for non-model plants.

6.1.2 Over-representation analysis (ORA)

In the past decade, researchers have developed a large number of knowledge bases to help to understand the transcriptome at functional level. The knowledge describes---using the standardized nomenclature of GO terms---the biological processes, components, and molecular functions in which individual genes and proteins are known to be involved in, as well as---using the not-so-standardized nomenclature of biological pathways---how and where gene products interact with each other. This allows the analysis of RNA-seq data at the functional level.

Using GO term and pathway datasets, some people have tried to identify active pathways that differ between two tissues/conditions based on a list of differentially expressed genes, in an approach generally known as over-representation analysis (ORA) [72]. However, there are several limitations in this approach: a) the gene list is selected by some statistical measurements (e.g. fold change, t-test); sometimes, the power of these measurements is reduced by sample size; b) just significant differential genes are selected, which may lose some information of other relevant genes; c) each gene is treated equally and assumed to be independent; d) each pathway and GO term is assumed independent of other pathways and GO terms, which is not true; and, most importantly, e) a slight change in the threshold of the test statistic can lead to a total change in the ORA outcome, rendering its conclusions rather unstable.

6.1.3 Direct-group Analysis

In order to deal with the limitations of ORA, some investigators have tried to consider the distribution of the pathway genes in the entire list of genes, and assign some functional class scores (FCS) to different GO terms or pathways. For this type of methods, a gene-level statistic is computed first using molecular measurements from an experiment, such Pearson correlation, ANOVA [23], t-test [73] and Z-score [22]. The gene-level statistics for all genes in a pathway/GO term are aggregated into a single pathway/GO-level statistic, and compared with a null distribution obtained from random gene sets of the same size as the reference gene set (i.e. the pathway or genes belonging to the GO term) being studied. The pathway-level statistics used by current approaches include sum, mean, median of gene-level statistic [182], the Wilcoxon rank sum [74] and the Kolmogorov-Smirnov-like statistic in GSEA [21]. Although FCS-type methods are an improvement over ORA, they still have several limitations. First, similar to ORA, they also analyze each pathway/GO term independently. Second, when the pathway contains too many non-causal genes, the statistical score can be largely affected. These methods are more likely to identify pathways that contain a sufficiently large proportion of disease-related genes, but pathways that contain only a few phenotype-related genes may be missed [24].

Further, some investigators attempted to incorporate some pathway topology information into the methods above. For example, Rahnenfuhrer et al. proposed ScorePAGE, which computes the similarity between each pair of genes in a pathway [183]. Then, the number of reactions needed to connect two genes in a given pathway is used to divide the pairwise similarities.

6.1.4 Network-based Analysis

In order to address the limitations that arise from direct-group analysis, network-based methods identify a subset of genes that might be most relevant to a phenotype for each pathway. They break up a pathway into smaller parts, called ‘sub-networks’. Methods in this category include NEA [184], SNet [25] and PFSNet [24]. However, they still have some limitations. First, similar to other methods, they also consider each pathway independently. Second, for an experiment with very few samples, it is impossible to compute and estimate the P-value of the test statistic used by these methods. Third, these methods need a pathway database that has relevant large pathways; they do not work if some part of relevant pathway is missing or the relevant pathway is too small.

6.1.5 Model-based Analysis

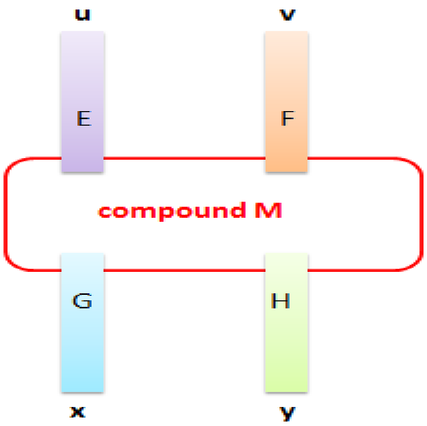
Model-based methods are a category of gene-set-based methods that attempt to learn parameters for a dynamic model of any given pathway using one phenotype [24]. Different methods may use different models for pathways, such as linear models in SRI [75] and Petri nets in GGEA [26]. For this type of methods, a major drawback is that parameters are difficult to estimate when developing different models for pathways.

6.2 Methods

These existing methods share a number of limitations that make them unsuitable for the investigation of plant secondary metabolites. In such cases, we are typically comparing a mutant to the wild type. This extremely small sample size presents a severe challenge to all the methods mentioned above, even to the extent of rendering them inapplicable. Moreover, for plants without a reference genome, their reference pathways are highly fragmented as there are many nodes in these pathways that we do not know what the corresponding genes/proteins are. This incompleteness directly affects the effectiveness of analysis methods that rely on pathway information. Lastly, almost all of the methods mentioned above consider each pathway independently. This does not seem reasonable in the context of metabolic pathways. In metabolic pathways, metabolites are produced and consumed. The amount of a metabolite that participates in two or more metabolic pathways has to be split among the two pathways [Figure 6.4]. Thus the activity (reflected as gene expression level) of the enzymes that process that metabolite should also be split among the two pathways. Analyzing gene expression in metabolic pathways without taking this into account potentially leads to more false positives and false negatives.

We take the issues above into consideration, and propose here a "Weighted Pathway" approach for gene expression analysis of plant secondary metabolic pathways based on next-generation sequencing data.

Figure 6.4 Model to deal with hub compound; Note: u,v,x,y denotes pathway; E,F,G,H denotes enzymes



We assume that the abundance level of a compound is roughly correlated to the gene-expression level of the enzymes catalyzing the production of that compound. After the gene-expression level of enzymes has been determined in one of the standard ways, the Weighted Pathway approach analyses metabolic pathways in three main steps. In the first step, the gene-expression level of each enzyme is adjusted taking into consideration the sharing of metabolites and enzymes across pathways. This produces the weighted pathways. In the second step, these weighted pathways are compared between the mutant and wild type and scored for significance. In the third step, which is an optional refinement, important sub-networks in weighted pathways are identified. These steps are presented in subsections below, along with the preparatory steps of preparing the plant metabolic pathway database used here and determining the initial gene expression values of enzymes.

6.2.1 Preparatory step 1: Database of plant metabolic pathway

The expanding demand for the production of food, feed, medicine, and biofuel from plants has prompted the sequencing of plant genomes and transcriptomes. To date, genome and mRNA sequences are available for a large number of plant species, and many more are under way. However, only a few genome-wide metabolic network reconstructions exist for plants. These include, but are not limited to, *Arabidopsis* and poplar maps inferred from KEGG reference maps [185], *Arabidopsis* and rice reactions and pathways inferred from reactome human maps [186], and a number of databases inferred from MetaCyc [187], such as AraCyc for *Arabidopsis* [188], RiceCyc for *rice* (http://pathway.gramene.org/gramene/ricecyc.shtml), MedicCyc for *Medicago truncatula* [189], LycoCyc for tomato (http://pathway.gramene.org/gramene/lycocyc.shtml) and ChlamyCyc for *Chlamydomonas reinhardtii* [190]. However, the lack of consistency in annotation standards and the lack of comparable quality in validation and curation hinder researchers seeking to meaningfully compare the metabolic networks of individual species housed in different metabolic databases.

PlantCyc [191] is a comprehensive plant metabolic pathway database, which is created to collect metabolic networks from other databases related to plants. They have already unified the format and definition for each plant. Hence, we just need to extract all the pathways from different plants in PlantCyc [191] and remove redundancy between different plants. The resulting database is used as our reference metabolic pathway in downstream analysis, as discussed later.

6.2.2 Preparatory step 2: Calculation of enzyme gene expression level

We assume next-generation sequencing of the transcriptome of the mutant and wild type has been performed. Then, in our study here, after the de novo assembly of sequence reads into transcripts, RSEM (RNA-seq by Expectation-Maximization) is used to estimate the abundance of assembled transcripts [192]. Each transcript is then mapped to an enzyme in our pathway database by homology search. If multiple transcripts are mapped to the same enzyme, the sum of the transcripts' abundance levels is used to represent the expression level of that enzyme. We call this value the “absolute expression level” of that enzyme. If we cannot find homologs in our de novo assembled transcripts for some enzymes in a reference pathway, their absolute expression levels are set to zero.

This preparatory step can be skipped or modified. For example, when gene expression values are directly supplied as input, we can simply use these gene expression values as the absolute expression levels of the enzymes.

We acknowledge that the transcript abundance level and protein abundance level are not always tightly correlated. Nevertheless, many people have found that transcriptome analysis has similar results to proteome analysis [193]. Hence, the transcript abundance level is used as the absolute expression level of enzymes as described above.

6.2.3 Main step 1: Relative gene expression level of enzyme

In contrast to the absolute expression level of an enzyme, the “relative expression level” of an enzyme is defined later in this section with respect to a pathway, and is intended to reflect the amount of activity of the enzyme that contributes to that pathway. For an enzyme that participates in multiple pathways, the sum of its relative expression levels with respect to these pathways is equal to its absolute expression level.

In order to make the definition, let us first pay attention to hub compounds which link multiple pathways in the metabolic pathway database. Without loss of generality, let us consider the example in Figure 6.4. Let us assume that, over a fixed unit of time, the metabolite M is produced in pathways u and v (can be more than 2), with amounts P(M,u) and P(M,v), and is consumed in pathways x and y (also can be more than 2), with amounts C(M,x) and C(M,y). Suppose the plant is in a steady state; i.e., the production and consumption of metabolites is in equilibrium. Thus:

$$P(M,u) + P(M,v) = C(M,x) + C(M,y)$$

Rearranging the above equation, we get:

$$C(M,x) = \frac{P(M,u) + P(M,v)}{C(M,x) + C(M,y)} \times C(M,x) \qquad equation\ 1$$

$$P(M,u) = \frac{C(M,x) + C(M,y)}{P(M,u) + P(M,v)} \times P(M,u) \qquad equation\ 2$$

Since we do not have data from direct measurement of the abundance of metabolites, the value of P(M,_) and C(M,_) are unknown. However, using some initial estimates of the abundance of metabolites in the pathways, we could get better estimates. We proceed in stages (the outer loop, which estimates the relative expression level of enzymes in specific pathways) and in rounds (the inner loop, which estimates the production and consumption level of metabolites in specific pathways):

$$C(M,x,i+1,h) = \frac{P(M,u,i,h) + P(M,v,i,h)}{C(M,x,i,h) + C(M,y,i,h)} \times C(M,x,i,h)$$

$$P(M,u,i+1,h) = \frac{C(M,x,i,h) + C(M,y,i,h)}{P(M,u,i,h) + P(M,v,i,h)} \times P(M,u,i,h)$$

The index i is used to indicate the estimates of the production and consumption levels of metabolites in the pathways at round i. The index h is used to indicate the estimates of the initial relative expression level of enzymes that produce or consume the metabolites in the pathways at stage h.

At each stage h, if we iterate sufficiently long and these estimates converge, we obtain:

$$C(M, x, k+1, h) = C(M, x, k, h)$$

$$C(M, y, k+1, h) = C(M, y, k, h)$$

$$P(M, u, k+1, h) = P(M, u, k, h)$$

$$P(M, v, k+1, h) = P(M, v, k, h)$$

We denote the index value i at convergence in stage h by i'. Therefore:

$$C(M, x, i', h) = \frac{P(M, u, i', h) + P(M, v, i', h)}{C(M, x, i', h) + C(M, y, i', h)} \times C(M, x, i', h)$$

$$P(M, u, i', h) = \frac{C(M, x, i', h) + C(M, y, i', h)}{P(M, u, i', h) + P(M, v, i', h)} \times P(M, u, i', h)$$

Hence, we can obtain the production and consumption value of M in the pathways u, v, x and y based on the estimates of the relative expression level of enzymes in the pathways in stage h. We call this procedure "adjust pathway", which tries to adjust the production/consumption level of metabolites in the pathways.

Since we do not directly measure the abundance level of metabolites, we estimate these values based on the relative expression level of the corresponding enzymes, which produce and consume the metabolites. Thus, the initial values are:

$$P(M, u, 0, h) = R(E, u, h)$$
$$P(M, v, 0, h) = R(F, v, h)$$
$$C(M, x, 0, h) = R(G, y, h)$$
$$C(M, y, 0, h) = R(H, y, h)$$

Here, R(_,_,h) denotes the stage-h estimates of the relative expression level of the corresponding enzymes in the specific pathways that produce or consume the metabolite M. Hence E and F are the enzymes that produce M in pathways u and v respectively, and G and H are the enzymes that consume M in pathways x and y respectively.

After convergence at round i’ in stage h, we can make the stage h+1 estimate of how much of the expression of an enzyme, for example E in Figure 6.4, that produces M contributes to a pathway x:

$$R(E,x,h+1) = A(E) \times \frac{C(M,x,i',h)}{C(M,x,i',h) + C(M,y,i',h)}$$

Here, A(E) is the absolute expression of an enzyme E that produces the metabolite M. Also, we initialize R(E,x,0)=A(E). This estimate of R(E,x,h+1) assumes, not unreasonably, that the production of the metabolite M by multiple enzymes are pooled before being consumed. The value R(E,x,h) is called the relative expression level of enzyme E in pathway x at stage h. We call this estimation procedure “split pathway”, which tries to split the absolute expression level of an enzyme into the pathways it is involved in.

However, in the whole pathway database, there are some specific metabolites, which have no producer or no consumer. In order to deal with these metabolites at the boundary of a pathway, we first merge all the pathways into one big pathway. Then some artificial start enzymes and end enzymes are added into pathways for these boundary compounds. The absolute expression level for these artificial enzymes is set to the sum of the absolute expression level of all the enzymes which produce/consume the respective compounds.

Each procedure---adjust pathway and split pathway---is run iteratively on hub compounds and enzymes that produce or consume hub compounds. During iteration, if the percentage change at each and every enzyme between two successive rounds is small enough [<5%], the iteration process stops. Note that enzymes that produce or consume non-hub compounds are not touched by this iteration process; thus their relative expression levels are equal to their absolute expression levels. We use the term “weighted pathway” to refer to a pathway annotated with the relative expression level of enzymes with respect to this pathway.

6.2.4 Main step 2: Identifying significant pathways

The overall expression level O(P, S) of a given weighted pathway P, in a given sample S, can be defined based on the relative expression level of enzymes with respect to that pathway. A simple choice is to set O(P,S) as the median or mean of the relative expression level of enzymes in pathway P in sample S. In this study, the mean is used.

There is no good applicable statistical method for determining which pathway P is significantly different in overall expression level between a mutant sample M and the wild-type sample W, due to this extremely small sample size of 2. A possible alternative is to consider the magnitude of the difference between the two overall expression levels, |O(P, M) – O(P, W)|. However, in this case, a small percentage difference between the two when O(P,M) is a high overall expression level can rank the pathway substantially higher than a large percentage difference when O(P,M) is a medium overall expression level. This is not reasonable. Another alternative is to consider the ratio between the two overall expression levels, O(P,M)/O(P,W). However, in this case, a small magnitude difference between the two when O(P,M) is a low overall expression level can still result in a high ratio. This is also problematic.

So we propose a practical compromise. We compute the overall expression level O(Pi, M) of each pathway Pi in the mutant sample M, and determine the mean μ_M and standard deviation σ_M of these overall expression levels. All pathways P_i with O(Pi,M) exceeding some threshold τ_M (in this study, $\tau_M = \mu_M$, but other thresholds e.g., $\tau_M = \mu_M + 2\sigma_M$ can be used) are kept as candidates. Similarly, we also determine the mean μ_W and standard deviation σ_W of the overall expression level of pathways in the wild-type sample W, and all pathways P_i with O(Pi,W) exceeding some threshold τ_W (in this study, $\tau_W = \mu_W$, but other thresholds e.g., $\tau_W = \mu_W + 2\sigma_W$ or even $\tau_W = \tau_M$ can be used) are also kept as candidates. Note that τ_M and τ_W need not have the same value. For each candidate P_i, we compute its fold change between the two phenotypes as the greater of the two ratios O(Pi,W)/O(Pi,M) and O(Pi,M)/O(Pi,W). Note that if the absolute/relative expression levels are in log base 2, fold change should be computed as $|O(P_i,W) – O(P_i,M)|$ or as $2^{|O(Pi,W) – O(Pi,M)|}$, depending on whether one prefers to think in log units. Candidate pathways are then ranked based on this change value. That is, we consider only pathways that have high overall expression level in either the mutant or the wild type, and rank them based on fold change.

6.2.5 Main step 3: Extracting sub-networks

A further refinement for pinpointing a more specific part of a weighted pathway that is likely to cause the difference between the two phenotypes is to generate and consider sub-networks of each weighted pathways. Given any enzyme E in a weighted pathway P, we generate the sub-network E_P by taking E and all the enzymes and metabolites that are down-stream of E in P to be the sub-network, and letting the enzymes in E_P inherit their relative expression levels (with respect to the sub-network) in the mutant and wild-type samples from P. We keep only those sub-networks having at least three enzymes (some other threshold is possible, but we use this threshold because a larger threshold would disqualify at least half the pathways in the database) and whose overall expression levels in the mutant sample M exceed τ_M or whose overall expression levels in the wild-type sample W exceed τ_W. We call these the candidate sub-networks.

A sub-network E_P in P is said to be an "ancestor" of another sub-network E_P' in P if, and only if, E_P is a subset of E_P'. In this case, we also say E_P' is a "descendant" of E_P. We consider a candidate sub-network E_P in P to be interesting if, and only if, the ratio of its overall expression levels between the two phenotypes is greater than, or equal to, that of every one of its ancestors and descendants in P that is also a candidate sub-network. Such an interesting sub-network E_P basically suggests that the enzyme E and its down-stream effects form the part in P that shows the biggest difference between the two phenotypes. Interesting sub-networks from all the pathways are ranked based on fold change.

6.3 Results

6.3.1 Plant metabolic pathway database

After removing redundancy from different plant metabolisms in PlantCyc (June 2013,lastest version) [191], 879 pathways, 3,455 compounds and 3532 reactions are used for our final plant metabolism pathway database. These numbers are a little different from the statistics from PlantCyc. This may be because they include the latest pathways submitted by other people, which have not yet been included in latest backbone database. Compared to other plant metabolism pathway databases, this one is much more comprehensive not only in the number of pathways captured, but also in the number of reactions. It also contains many more pathways, compared to KEGG [185], which collects pathways from all organisms [Table 6.1].

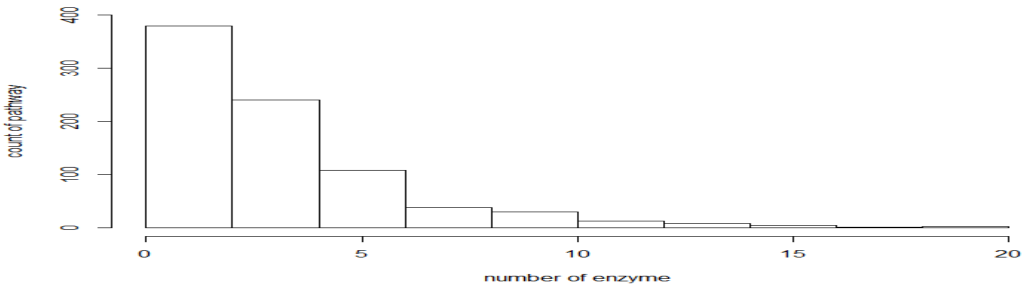
Table 6.1 Statistic information for different pathway database

	Pathways	Enzymes	Reactions	Compounds
PlantCyc	1,050	188,798	5,332	4,410
AraCyc	597	9,041	3,490	2,613
BarleyCyc	465	7,572	2,901	2,135
BrachypodiumCyc	473	8,802	2,915	2,128
CassavaCyc	491	10,007	3,058	2,232
ChineseCabbageCyc	499	10,976	3,104	2,270
ChlamyCyc	349	3,330	2,263	1,514
CornCyc	508	14,818	2,958	2,271
GrapeCyc	479	7,572	3,015	2,229
MossCyc	416	7,805	2,713	1,901
OryzaCyc	482	15,677	3,000	2,226
PapayaCyc	481	5,714	2,999	2,220
PoplarCyc	505	20,822	3,124	2,295
SelaginellaCyc	421	6,462	2,737	1,987
SetariaCyc	477	10,214	2,942	2,145
SorghumBicolorCyc	480	8,630	2,939	2,141
SoyCyc	520	20,317	3,105	2,273
SwitchgrassCyc	479	17,319	2,985	2,184

KEGG	455 [172 relate to metabolism]	6,166	9,485	17,150

However, from the length distribution, we can see that our pathways tend to be short ones. Nearly half of the pathways in our database have just 3 enzymes [Figure 6.5]. For this type of pathways, network-based pathway analysis methods---which attempt to find enriched sub-networks in a longer pathway---are not suitable.

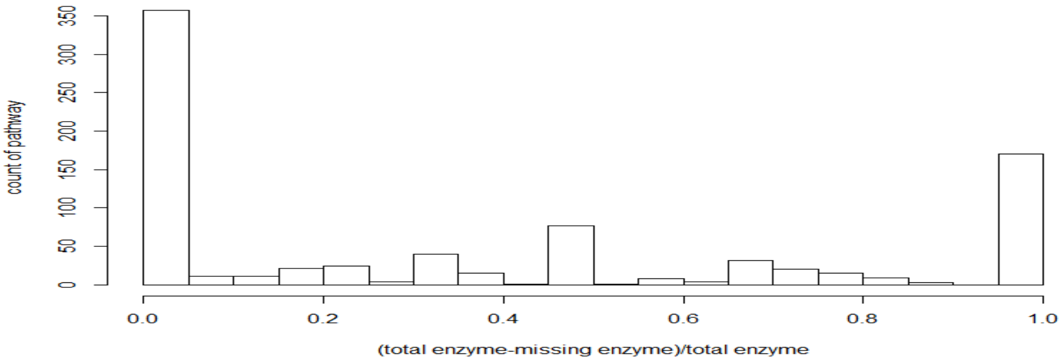
Figure 6.5 Histogram of length of pathways in our database



For the plants we are studying, there are no reference genomes and very little sequencing information in public databases. The most common method for enzyme annotation is still just by homology search. However, the effect of homology search depends on the completeness of the enzyme database. The number of enzymes for non-model plants in the database is still limited, which makes it is a challenge to map de novo assembled transcripts for enzymes in the database. For the pathways in our database, many pathways still have some missing enzymes [Figure 6.6]. However, it does not mean that these missing enzymes are really missing in our assembled unigenes. It may be just because we cannot find them in our assembled results.

Therefore, it raises some challenges in applying traditional pathway-based gene expression profile analysis methods to them. If one or two enzymes are missing in these pathways, most traditional methods become inapplicable to our dataset. This is because most of them use correlation between different enzymes in the same pathway to score this pathway. If one pathway is left with just one or two non-missing enzymes, these correlation scores are meaningless.

Figure 6.6 Histogram for missing enzyme ratio in our pathway database



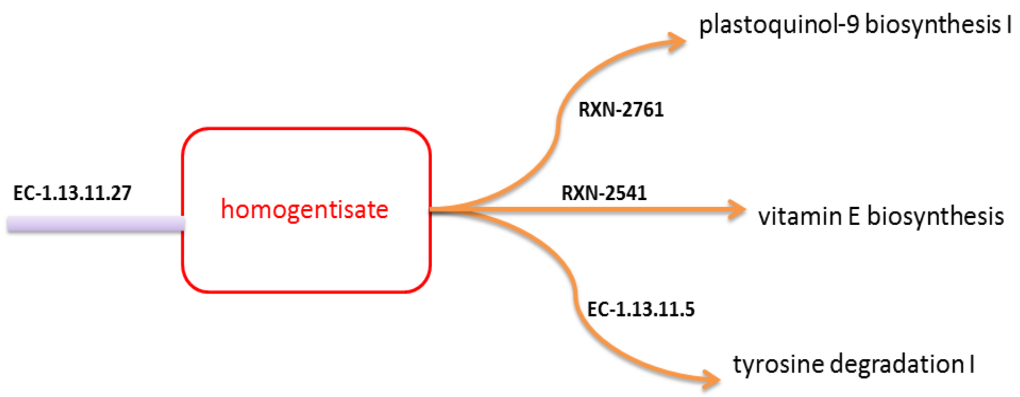
6.3.2 Validity of weighted pathway approach

In order to verify the correctness of weighted pathway approach, two public datasets from *Arabidopsis thaliana* are used: *VTE2* mutant and *SID2* mutant.

6.3.2.1 *VTE2* mutant

The enzyme EC-1.13.11.27 catalyzes the production of homogentisate, which is the substrate for enzymes RXN-2761 (for plastoquinol-9 biosynthesis I), RXN-2541 (for vitamin E biosynthesis), and EC-1.13.11.5 (for tyrosine degradation I); see Figure 6.7. Our first validation dataset is a public dataset (GSE53990) of the VTE2 mutant, in which the enzyme RXN-2541 is mutant. According to the experiments of Michel et al. [194], the level for vitamin E (tocopherols) is greatly reduced in the VTE2 mutant, compared to the wild type. Also, the level for carotenoids (plastoquinol-9 biosynthesis I is precursor) is elevated.

Figure 6.7 Model for VTE2 mutant in *Arabidopsis*



We apply Weighted Pathway to analyse this dataset. As a preparatory step, RMA (Robust Multiple-Array Average) [195] is applied to obtain the absolute expression level (log base 2) for each gene. Then the relative expression level for each gene is computed as described in the first step of Weighted Pathway. We can already see a clear difference in the relative expression levels of the enzyme EC-1.13.11.27 in the three pathways (vitamin E biosynthesis, tyrosine degradation I, plastoquinol-9 biosynthesis I) between VTE2 and the wild type (WT), whereas there is no difference in the absolute expression levels of the enzyme EC-1.13.11.27 in VTE2 and the wild type in these pathways; see Table 6.2.

Table 6.2 Expression level for enzyme EC-1.13.11.27. WT and VTE2: denote expression level using absolute expression level; WT_weighted and VTE2_weighted: denote using our weighted pathway model

EC-1.13.11.27	WT	VTE2	WT_weighted	VTE2_weighted
plastoquinol-9 biosynthesis I	7.588548	7.58976	7.315344	7.443480154
vitamin E biosynthesis	7.588548	7.58976	4.166117	2.071157721
tyrosine degradation I	7.588548	7.58976	3.931097	3.891596291

We also computed the overall expression level of each pathway based on the mean of relative expression levels of enzymes in the pathway, as described in the second step of Weighted Pathway (μ_W=4.97 for WT and μ_M=4.93 for VTE2 mutant). We see that the vitamin E biosynthesis pathway has a 1.54-fold (= 0.619 in units of log base 2) reduction in overall expression in VTE2 compared to the wild type [Table 6.3]. Considering only candidate pathways having overall expression level greater than mean in either WT or VTE2 mutant, we find that rank of vitamin E biosynthesis pathway is 9 [Table 6.4]. If we exclude pathways whose size is smaller than 3, the rank for this pathway is slightly improved [Table 6.4]. In addition, we see that the plastoquinol-9 biosynthesis I pathway has a slight 1.04-fold (= 0.056 in units of log base 2) increase in overall expression in VTE2 compared to wild type [Table 6.3]. Similarly, also considering only candidate pathways having expression level greater than mean in either WT or VTE2 mutant, rank of plastoquinol-9 biosynthesis pathway is 229 [Table 6.4].

Table 6.3 Mean value for different pathway WT and VTE2 denotes mean value using absolute expression level; WT_weighted and VTE2_weighted denotes the mean value using our weighted pathway model

	WT	VTE2	FC	WT_weighted	VTE2_weighted	FC
vitamin E biosynthesis	7.940	7.633	-0.307	6.257	5.637	-0.619
tyrosine degradation I	6.059	5.993	-0.066	5.786	5.340	-0.406
plastoquinol-9 biosynthesis I	9.008	9.020	0.012	7.457	7.513	0.056

After applying the third step in Weighted Pathway, the rank does not show a big difference [Table 6.4]. This is because the pathways in the database are very small and, hence, not many significant sub-networks are identified.

Table 6.4 Rank for different pathways based on relative expression level for VTE2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than the mean in the wild type or mutant; rank (mean & size>3) denotes rank using pathways having relative expression level more than mean in wild type or mutant and size should be more than 3; rank (sub-network) denotes rank using sub-networks.

	rank(all)	rank(>mean)	rank (mean&size>=3)	rank (sub-network)
vitamin E biosynthesis	11	9	7	10
tyrosine degradation I	29	23	13	28
plastoquinol-9 biosynthesis	314	229	128	247

In contrast, if we compute the overall expression level of a pathway as the mean of the absolute expression levels of enzymes in that pathway, we would see a 1.24-fold (= 0.307 in units of log base 2) reduction in the overall expressional level of the vitamin E biosynthesis pathway and a 1.01-fold (= 0.012 in units of log base 2) increase in the overall expression level of the plastoquinol-9 pathway [Table 6.3]. Considering pathways having expression level greater than mean in either WT or VTE2 based on absolute expression level, we find that rank for vitamin E biosynthesis pathway is 30, and rank for plastoquinol-9 pathway is 328 [Table 6.5]. It is clear that Weighted Pathway has more clearly identified the experimental observations of Michel et al. [Figure 6.8] [194].

Table 6.5 Rank for different pathways based on absolute expression level for VTE2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than the mean in the wild type or mutant; rank (sub-network) denotes rank using sub-networks.

	rank(all)	rank(>mean)	rank (sub-network)
vitamin E biosynthesis	34	30	38
tyrosine degradation I	330	250	321
plastoquinol-9 biosynthesis	524	328	460

Although there is no direct evidence showing the decrease of tyrosine degradation I pathway under VTE2 mutant, several groups have reported that there is a correlation between vitamin E biosynthesis and tyrosine aminotransferase (the first enzyme in the tyrosine degradation pathway) [196-198]. Hence, it is also consistent with our results, which shows relatively large change in the tyrosine degradation pathway.

Figure 6.8 Vitamin E level for wild type and VTE2 mutant in *Arabidopsis* [194]

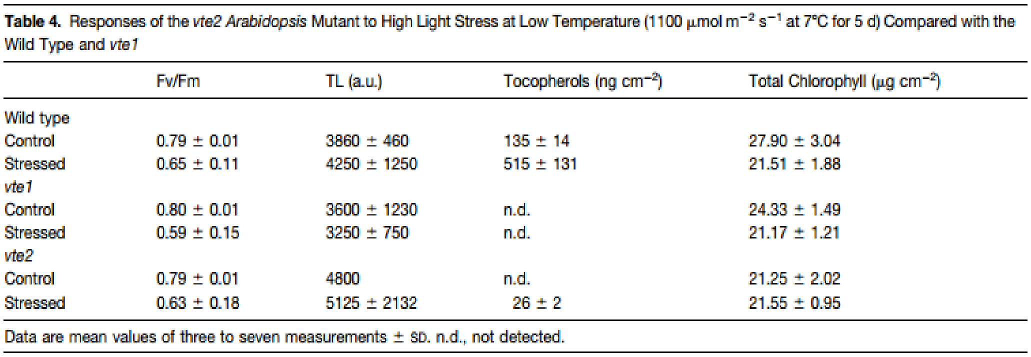


Table 4. Responses of the *vte2 Arabidopsis* Mutant to High Light Stress at Low Temperature (1100 μmol m^{-2} s^{-1} at 7°C for 5 d) Compared with the Wild Type and *vte1*

	Fv/Fm	TL (a.u.)	Tocopherols (ng cm^{-2})	Total Chlorophyll (μg cm^{-2})
Wild type				
Control	0.79 ± 0.01	3860 ± 460	135 ± 14	27.90 ± 3.04
Stressed	0.65 ± 0.11	4250 ± 1250	515 ± 131	21.51 ± 1.88
vte1				
Control	0.80 ± 0.01	3600 ± 1230	n.d.	24.33 ± 1.49
Stressed	0.59 ± 0.15	3250 ± 750	n.d.	21.17 ± 1.21
vte2				
Control	0.79 ± 0.01	4800	n.d.	21.25 ± 2.02
Stressed	0.63 ± 0.18	5125 ± 2132	26 ± 2	21.55 ± 0.95

Data are mean values of three to seven measurements ± SD. n.d., not detected.

6.3.2.2 *SID2* mutant

The enzyme EC-4.2.3.5 catalyzes the production of chorismate, which is the substrate for enzymes EC-5.4.4.2 (for salicylate biosynthesis I and 1,4-dihydroxy-2-naphthoate biosynthesis II), EC-2.6.1.85 (for 4-aminobenzoate biosynthesis and tetrahydrofolate biosynthesis II), EC-4.1.3.27 (for tryptophan biosynthesis) and EC-5.4.99.5 (for phenylalanine biosynthesis II and tyrosine biosynthesis); see Figure 6.11. Our second validation dataset is a public dataset (GSE25489) of the SID2 mutant, in which the enzyme EC-5.4.4.2 (ICS/SID2) is mutant. In *Arabidopsis thaliana*, systemic acquired resistance against pathogens has been associated with the accumulation of salicylic acid (SA) [199]. Garcion et al. have demonstrated the function and localization of ICS involved in SA biosynthesis [Figure 6.9] [200]. Furthermore, reduction was also observed for phylloquinone production by the 1,4-dihydroxy-2-naphthoate biosynthesis II pathway [Figure 6.9] [200].

Figure 6.9 Functional roles of ICS. phylloquinone (B) and SA accumulation following UV induction (C) [200]

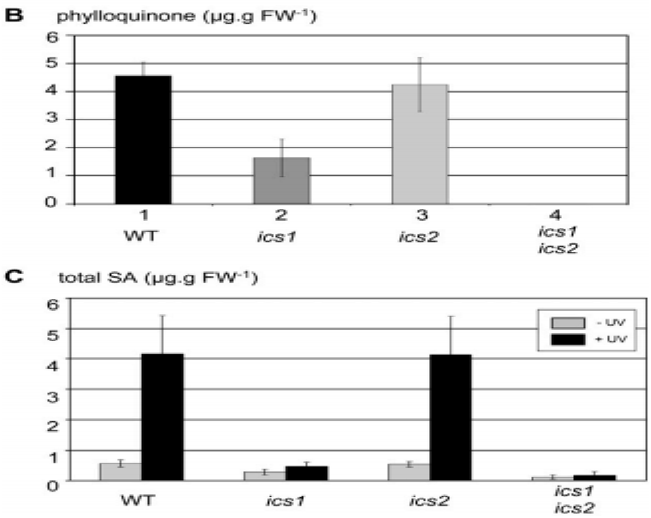
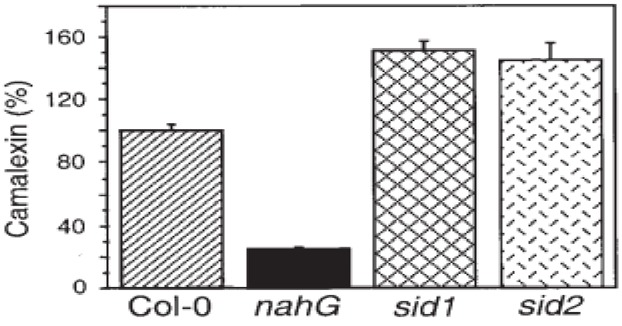
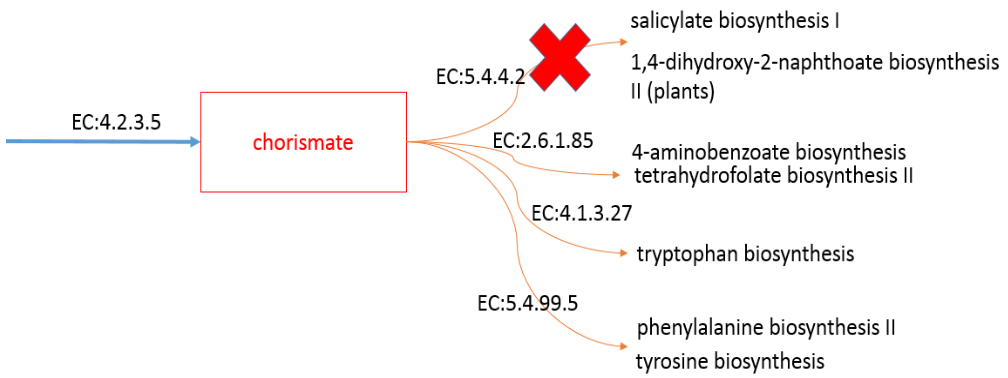


Figure 6.10 Accumulation of Camalexin in Leaves of *Arabidopsis* Col-0 Plants, NahG Plants (control), and sid (ICS) Mutant [199].



In addition, another group have demonstrated that camalexin (derived from tryptophan biosynthesis pathway) levels in SID2 mutant were higher compared to wild type plants [Figure 6.10] [199, 201]. Camalexin plays a role in resistance against pathogens, as does SA. Hence, they hypothesize that in *Arabidopsis* *thaliana*, there may be several independent ways leading to disease resistance. After checking the whole pathway database, we find that all of these three pathways share the same intermediate compound chorismate with several other pathways. In other words, after silencing of SID2/ICS (EC: 5.4.4.2), the whole flow is shifted between these pathways [Figure 6.11].

Figure 6.11 pathway model for ICS (SID2) mutant


EC:4.2.3.5
chorismate
EC:5.4.4.2
salicylate biosynthesis I
1,4-dihydroxy-2-naphthoate biosynthesis II (plants)
EC:2.6.1.85
4-aminobenzoate biosynthesis
tetrahydrofolate biosynthesis II
EC:4.1.3.27
tryptophan biosynthesis
EC:5.4.99.5
phenylalanine biosynthesis II
tyrosine biosynthesis

Similar to the VTE2 dataset, we also apply Weighted Pathway to analyze this dataset. Then, relative gene expression level for each enzyme is computed for each pathway as described in the first step of Weighted Pathway. We can also see a clear difference in the relative expression levels of the enzyme EC-4.2.3.5 in the three pathways (salicylate biosynthesis I, 1,4-dihydroxy-2-naphthoate biosynthesis II, tryptophan biosynthesis) between SID2 and the wild type (WT), whereas there is no difference in the absolute expression levels of the enzyme EC-4.2.3.5 in SID2 and the wild type in these pathways; see Table 6.6.

Table 6.6 Expression level for enzyme EC-4.2.3.5 in WT and ICS mutant. WT and Mutant denote the absolute expression level. WT_weighted and Mutant_weighted denote the relative expression level by our weighted pathway model.

EC-4.2.3.5	WT	Mutant	WT_weighted	Mutant_weighted
salicylate biosynthesis I	11.60815	11.83651	8.384705672	5.367518991
1,4-dihydroxy-2-naphthoate biosynthesis II	11.60815	11.83651	2.821442017	0.1265

tetrahydrofolate biosynthesis II	11.60815	11.83651	2.804929406	3.211543938
4-aminobenzoate biosynthesis	11.60815	11.83651	2.806005723	3.256979408
tryptophan biosynthesis	11.60815	11.83651	10.54118343	11.13955004
phenylalanine biosynthesis II	11.60815	11.83651	8.986942389	9.268036134
tyrosine biosynthesis II	11.60815	11.83651	9.237783499	9.496374756

We also compute the overall expression level of each pathway based on the mean relative expression levels of enzymes in the pathway, as described in the second step of Weighted Pathway (with a μ_W=5.55 for WT and μ_M=5.51 for SID2 mutant). We observe that the salicylate biosynthesis pathway has a 4.02-fold (= 2.007 in units of log base 2) reduction in overall expression in SID2 compared to wild type [Table 6.7]. We also see that the 1,4-dihydroxy-2-naphthoate biosynthesis II pathway has a 1.68-fold (= 0.75 in units of log base 2) decrease in overall expression in SID2 compared to wild type [Table 6.7]. Considering only candidate pathways having expression level greater than mean in either WT or SID2 mutant, the ranks of these two pathways are 1 and 9 respectively, the rank of the tryptophan biosynthesis pathway is 22 [Table 6.8]. After applying the third step in Weighted Pathway, the ranks are not changed much [Table 6.8]. Again, this is due to the small size of our pathways.

Table 6.7 Mean value for different pathway. WT and ICS denotes mean value using absolute expression. WT_weighted and ICS_weighted denote mean value using relative expression.

pathway level	WT	ICS	FC	WT_weighted	ICS_weighted	FC
salicylate biosynthesis I	10.45	9.53	**-0.926**	9.37503	7.3673	**-2.007**
1,4-dihydroxy-2-naphthoate biosynthesis II (plants)	7.57	7.22	**-0.349**	5.75	5.003964	**-0.75**
tetrahydrofolate biosynthesis II	8.45	8.58	0.13	6.88	7.067	0.187
4-aminobenzoate biosynthesis	8.11	8.219	0.103	0.93	1.08566	0.15
tryptophan biosynthesis	10.7	11.09	**0.388**	10.57	11.016	**0.44**

phenylalanine biosynthesis II	10.5	10.7	0.126	9.709	9.795	0.086
tyrosine biosynthesis II	9.894	10.04	0.149	9.121	9.223	0.101

Table 6.8 Rank for different pathways based on relative expression level for SID2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than mean in WT or mutant; rank (mean & size>3

	rank (all)	rank(mean)	Rank (mean & size>=3)	rank(sub-network)
salicylate biosynthesis I	1	1	1	1
1,4-dihydroxy-2-naphthoate biosynthesis	9	9	5	9
tetrahydrofolate biosynthesis II	113	100	46	129
4-aminobenzoate biosynthesis	144	X	X	161
tryptophan biosynthesis	24	22	11	24
phenylalanine biosynthesis II	242	196	103	258
tyrosine biosynthesis II	214	176	90	231

In contrast, if we compute the overall expression level of a pathway based on the mean of the absolute expression levels of enzymes in that pathway, we would see a 1.9-fold (= 0.926 in units of log base 2) reduction in the overall expressional level of the salicylate biosynthesis pathway and a 1.27-fold (= 0.349 in units of log base 2) decrease in the overall expression level of the 1,4-dihydroxy-2-naphthoate biosynthesis II pathway [Table 6.7]. Considering pathways having expression level greater than the mean in either WT or SID2 based on absolute expression level, we find that the rank for salicylate biosynthesis pathway is 2, and the rank for 1,4-dihydroxy-2-naphthoate biosynthesis II is 50 [Table 6.9]. It is clear that Weighted Pathway is able to more clearly identify the experimental observations of Garcion et al [200] and Schlaeppi et al [201].

Table 6.9 Rank for different pathways based on absolute expression level for SID2 mutant. rank (all) denotes rank using all the pathways; rank (>mean) denotes rank using pathways having relative expression level more than mean in WT or mutant; rank (sub-network)

	rank(all)	rank(>mean)	rank(sub-network)
salicylate biosynthesis I	2	2	2
1,4-dihydroxy-2-naphthoate biosynthesis	56	50	37
tetrahydrofolate biosynthesis II	110	96	125
4-aminobenzoate biosynthesis	257	199	271
tryptophan biosynthesis	41	38	34
phenylalanine biosynthesis II	220	180	232
tyrosine biosynthesis II	173	143	191

In summary, by these two simple examples, we can say that Weighted Pathway approach gives more reasonable results than the method using absolute expression level.

6.4 Conclusion

In this chapter, a weighted pathway model has been proposed to investigate different secondary metabolites for different plants. Different from previous analysis, the main advantage for our model is that we not only focus on the known pathways/compounds, but also try to predict the new functions/pathways for studied plants. We do not consider each pathway to be independent. Instead, two ideas---hub enzymes and hub compounds---are introduced into our model. From the verification results, we find that our weighted pathway approach is much more reasonable than traditional pathway analysis methods, which use absolute expression level.

We believe our weighted pathway approach will not only predict new functions/metabolites, but also provide more clues/ideas for future research about secondary metabolites. We will demonstrate this, in the next chapter, by applying Weighted Pathway to analyze mint secondary metabolism.

Chapter 7

APPLICATION ON SECONDARY METABOLISMS

7.1 Background

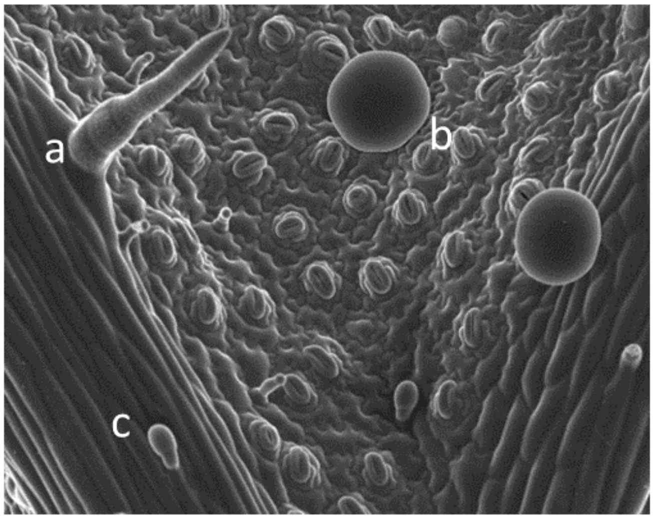
Plants produce enormous variety of specialized metabolites among which terpenes are the largest and most structurally varied class of natural products. Many of these terpenes are produced and stored in specialized secretory structures called glandular trichomes [Figure 7.1]. They are the main components of plant essential oils. These terpenes provide protection for plants against a variety of herbivores and pathogens [17, 168] and are also commercially quite valuable. But our knowledge about the development of secretory glandular trichomes and terpene production and its regulation is very limited, making it difficult to engineer these metabolic pathways.

Aromatic essential oil produced by *Mentha* species is the source of the best-known monoterpenes, menthol and carvone, which form the principal components of mint oil. They are extensively used in flavour and fragrance industries, pharmaceuticals and cosmetic products [202].

From the trichome of peppermint variety (*Mentha piperita*), 1,316 randomly-selected cDNA clones, or expressed sequence tags (ESTs) were produced, which led to the identification of many genes, enzymes and substrates involved in the main essential oil biosynthetic pathway [203, 204]. Given the technical limitations at their time of study, an EST approach would possibly identify only cDNAs which are abundant in trichome. A recent proteomic analysis of spearmint PGT identified 1,666 proteins of which 57 were predicted to be involved in secondary metabolisms [205]. But generation of sufficient genomic information with deep coverage is required to gain insights into the regulatory mechanism of terpene metabolism and glandular trichome development. This will promote successful engineering for improved yields or to develop mint as a platform for production of novel / altered terpenes. Mint is a well-suited plant for this as it is able to produce and store large amount of oils within trichome instead of exuding it on to the leaf surface. Storage within the trichomes also reduces the loss of volatile oils by emission into the atmosphere.

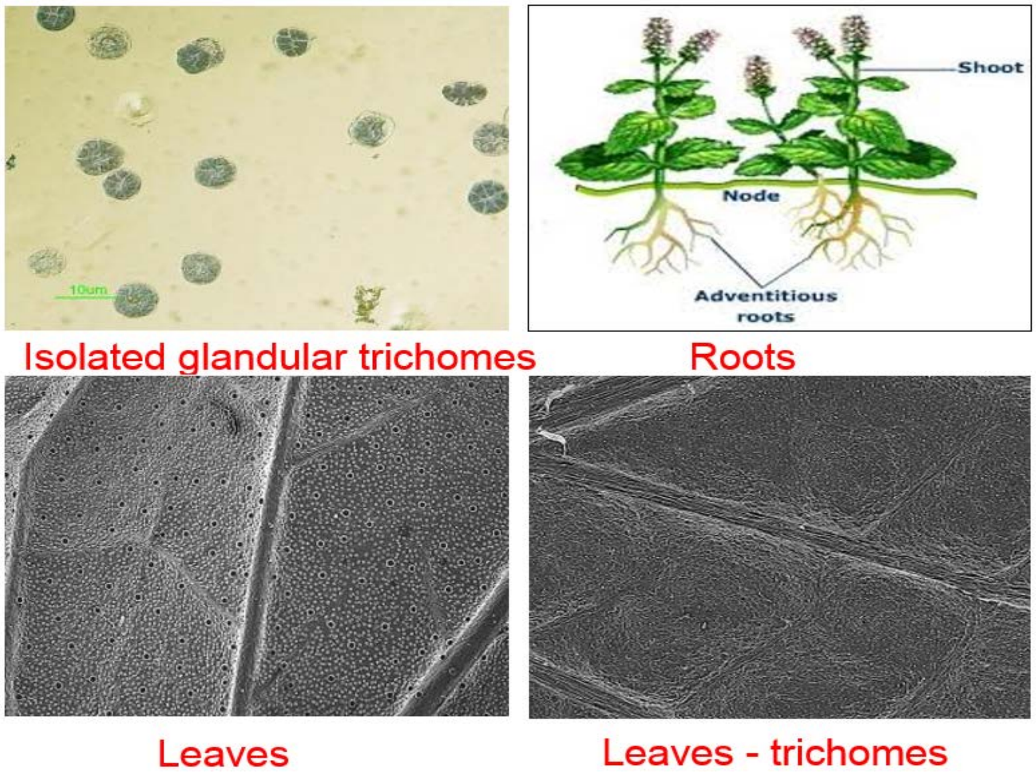
High-throughput RNA sequencing (RNA-Seq) has increasingly become the technology of choice to generate a comprehensive and quantitative profile of the gene expression pattern of a tissue. Here, we try to give a comparative analysis of RNA-seq transcriptome of different tissues of mint---namely trichome, leaf without trichome (leaf-trichome), and leaf. In addition, we are also interested in whether mint has the capacity to produce other secondary metabolites and whether it is possible to engineer other secondary metabolisms using mint as the platform. Hence, in this chapter, we use our weighted pathway approach developed in the last chapter to understand the metabolic capacity of mint.

Figure 7.1 Trichomes on spearmint leaf. a:Non glandular hairy trichome, b:Peltate glandular trichome (PGT), c: Capitate glandular trichome



7.2 Methods

Figure 7.2 The studied tissue for RNAseq strategy



7.2.1 RNA sequencing

For mint samples, RNA libraries for trichome tissues and other control tissues are prepared [Figure 7.2] and sequenced by Illumina following the manufacturer's instructions. Around 100 million reads of 101bp clean reads are generated from leaf, root, trichome and leaf without trichome (leaf-trichome) tissues, respectively [Table 7.1], achieving a higher coverage compared to previous EST databases [17].

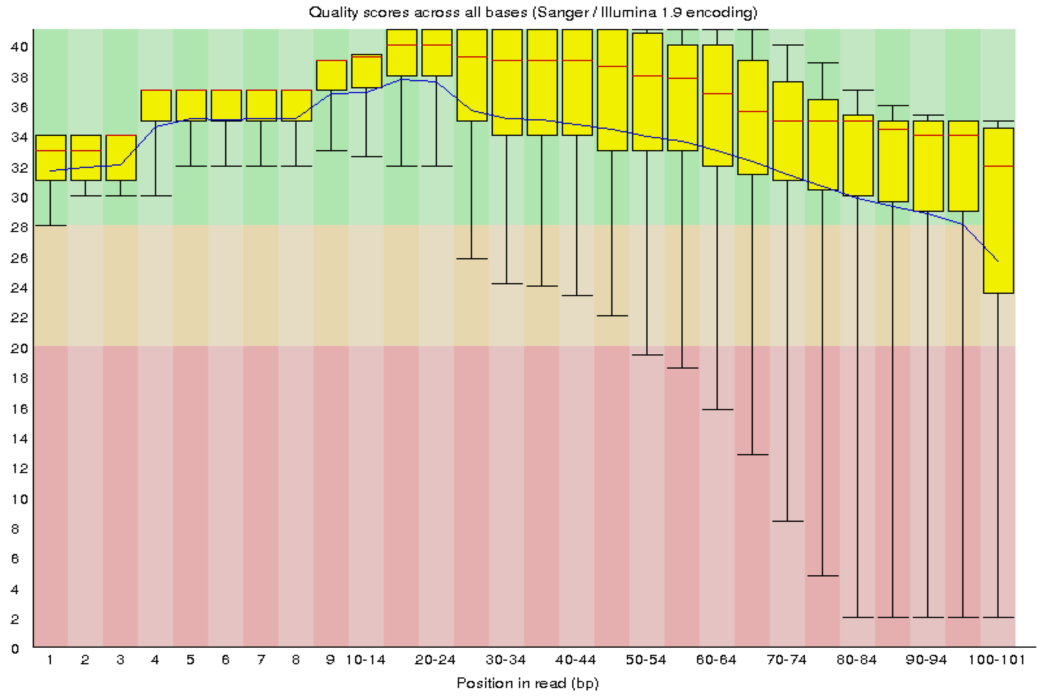
Table 7.1 Statistic for RNA seq results

organism	tissue	avg read size	raw reads	pair end
mint	leaf	101	115,404,986	yes
mint	root	101	91,153,220	yes

mint	Leaf-trichome	101	136,558,099	yes
mint	trichome	101	115,191,961	yes

As a first step, the quality for these sequencing reads is checked using the fastqc [http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc] tool box. If the raw reads have high quality [more than 70% bases having Quality score>=20, Figure 7.3], they are used in the next step. As these plants lack reference genome, only de novo assembly methods can be used. In the past several years, a lot of tools---such as SOAPdenovo [55], velvet [48], Oases [52] etc---have been developed for this purpose. Owing to the specific features for RNA assembly, such as alternative splicing, Trinity outperforms other methods, especially for RNA assembly [69]. Hence, it is used to perform the de novo assembly for our RNA sequencing results.

Figure 7.3 Quality control for RNA seq result (box plot for each position in read) x-axis: each base in read (bp) y-axis: quality score for each base/position (20: base accuracy is 99%, 30: base accuracy is 99.9%)



7.2.2 Weighted pathway analysis

As introduced in Chapter 6, after applying de novo assembly methods to assemble RNA-seq reads into transcripts, which represent different genes, transcriptome levels are mapped to enzymes as absolute expression levels. If we cannot find homologs in our de novo assembled transcripts for some enzymes, their absolute values are set to zero.

Then, the relative expression levels for all enzymes producing/consuming hub compounds in the pathways are computed using our weighted pathway approach proposed in Chapter 6. Using the resulting relative expression levels of all enzymes in a pathway, the overall expression level of the pathway is computed. By comparing overall expression level for each pathway between wild type and control, the enriched pathways are identified. We predict the potential new functions for our plants based on the enriched pathways.

7.3 Results

7.3.1 Results for RNA-seq

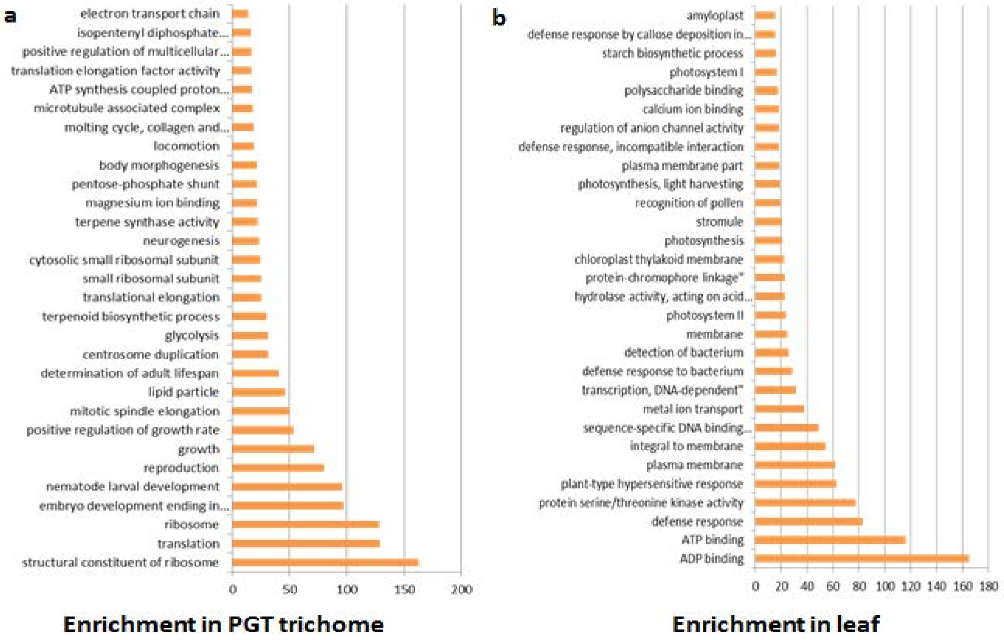
In total, more than 40,000 unigenes [a unigene is a hypothetical gene represented by a cluster of similar transcripts that are thought to be isoforms in a de-novo transcriptome assembly] have been assembled for all species [Table 7.2]. This is more than the typical number of genes for most known organisms [2, 97]. Some genes may be partial ones; some of them may be non-coding genes, like lincRNAs.

Table 7.2 Assembly results for the plant samples in our study

Species	# isoforms	# unigenes	total bases	N50	GC percentage
spearmint	87,480	40,587	101,396,693	1774	43.14%

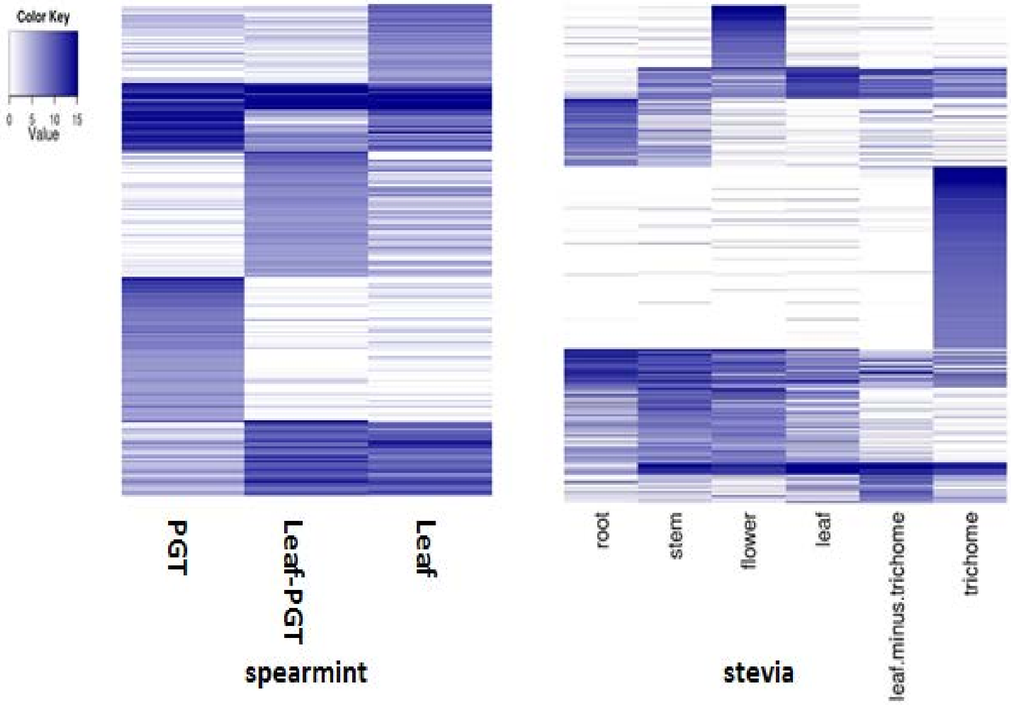
Functions of the unigenes are annotated based on sequence similarity to sequences in the public NR database [206]. At the same time, the protein sequence databases for *Arabidopsis*, *Vitis* and rice are also searched for homologs. For the mint dataset, among the 40,587 non-redundant unigenes, 27,025 (67%) have at least one hit in BLASTX search with E-value <= 1e-3. Functional classifications (GO term assignment) of all unigenes are done using Trinotate [69]. Then, the top 5000 up-regulated and down-regulated genes in trichome compared to leaf without trichome are selected for identifying enriched GO terms separately by hyper-geometric test.

Figure 7.4 Enrichment GO items by hypergeometric test. X-axis: log(1/p-value) a) Enrichment GO for trichome tissue of spearmint b) enrichment GO for leaf tissue of spearmint



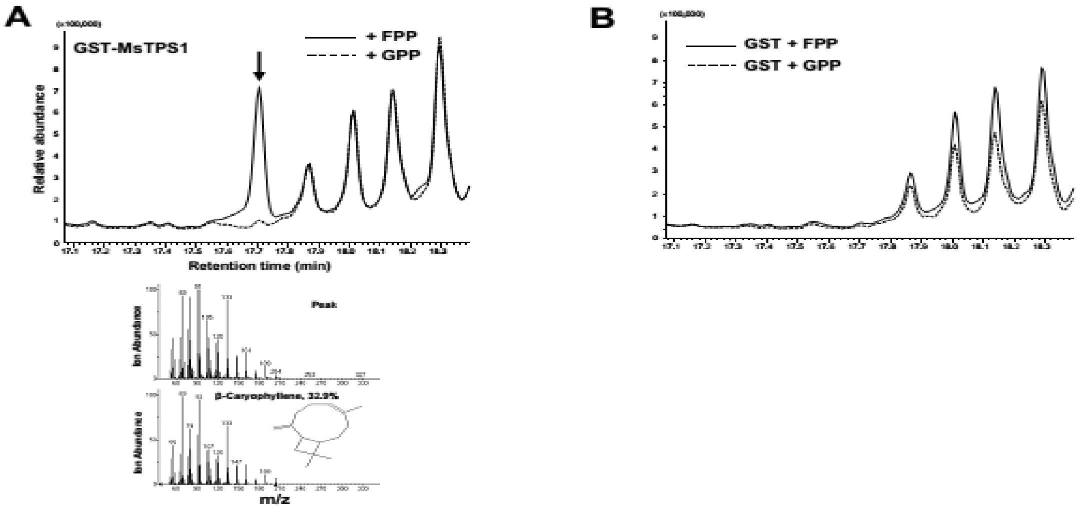
From the top 20 enriched GO terms for trichome and leaf tissue [Figure 7.4], we see that the photosynthesis-related GO terms are only enriched in leaf tissue; while terpene synthase-related GO terms are only enriched in trichome tissue. This is consistent with the fact that terpene are only produced in trichome of mint [207]. However, it is still unknown why mint trichome does not have photosynthesis-related GO function, which is a major source of energy.

Figure 7.5 Heatmap for different tissue in spearmint and stevia samples



In order to obtain the relative expression level for the assembled unigenes of each tissue, we first map reads onto them using bowtie [208]. RSEM [192] is then used for abundance estimation for the assembled transcripts to measure the expression level. From the heatmap shown in Figure 7.5, some specific patterns for trichome tissue in mint can be found. In other words, trichome tissue has some specific functions different from leaf and root. Among the specific patterns for trichome tissue, genes like P450, terpene biosynthesis, lipid transfer proteins (LTP) and interesting transcription factors like MYB, NAC are found, which may show good evidence for their potential function for the biosynthesis of specific secondary metabolites. An interesting finding is that we do not find any transcription factors (TFs) that matched major known trichome initiating TFs from *Arabidopsis* like TTG1, GL2,GL3 or Gl1 [209]. This may be the difference between glandular trichome for mint and stevia and non-glandular trichome in *Arabidopsis*. In addition, one of the terpene synthase genes (TPS) in mint which showed a trichome-specific pattern [Figure 7.5] and has no homolog in the NR [206] database, has now been functionally characterized as sesquiterpene synthase in our lab [Figure 7.6].

Figure 7.6 In vitro enzymatic assays of recombinant MsTPSs. GST-tagged MsTPS recombinant enzymes were purified by glutathione-based affinity chromatography and used for in vitro assays with GPP or FPP as substrate. The final products were analysed by GC-MS.



7.3.2 Results for weighted pathway approach

7.3.2.1 Enriched pathway for weighted pathway approach

As mentioned in the method part, after computing the relative expression level for each enzyme in the pathway database, we also compare the overall expression level of pathways by different measurements, such as mean, median, sum and FCS methods. We summarized the top 20 enriched pathways by our weighted pathway approach in Table 7.3. The detail explanation for these results is shown in sections below.

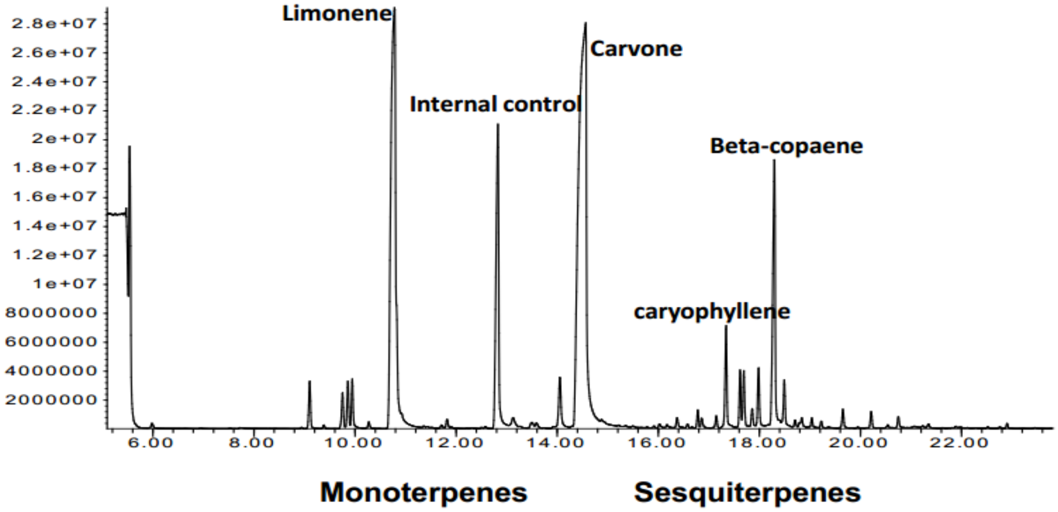
Table 7.3 Top 20 enrichment pathway for trichome and other tissue in mint by our weighted pathway model Where each row denotes a pathway; column (leaf, root, leaf-trichome, trichome) denotes the overall expression level for a pathway by mean value of the enzyme in the pathway; FC denotes fold change between trichome and leaf-trichome using mean overall value; median and sum denotes overall expression level for trichome tissue by median value and sum value of the enzymes in the pathway; Pearson denotes the score for a pathway by the average Pearson correlation among one pathway; scorePAGE denote the score computed by scorePAGE method [183]

pathway	root	leaf	leaf-trichome	trichome	FC	median	sum	pearson	scorePAGE
(4R)-carvone biosynthesis	58.1	592	128.6	5864	45.6	3879.3	17594	0.99	0.996
S-adenosyl-L-methionine biosynthesis	1741.3	880	802.7	3545	4.4	3545.1	3545.1	0.000	0.000
methylerythritol phosphate pathway	62.3	288	239.7	1674	7.0	1184.4	15064.4	0.929	0.396
adenine and adenosine salvage VI	713	463	410	1648	4	1648	1648	0	0
geranylgeranyldiphosphate biosynthesis	86.8	288	154.6	1395	9.0	1394.6	1394.6	0.00	0.000
menthol biosynthesis	111	221	325.8	961	3.0	638.7	8653	0.26	0.149
2'-deoxymugineic acid phytosiderophore biosynthesis	439.5	221	200.9	886	4.4	0.1	3545.3	-0.13	-0.033
free phenylpropanoid acid biosynthesis	151.3	63	33.6	851	25.4	851.9	851.9	0.00	0.000
trans, trans-farnesyl diphosphate biosynthesis	42.9	98	110.8	744	6.7	571.8	2232	0.78	0.789
pinobanksin biosynthesis	23.0	107	97.7	683	7.0	25.0	3416	-0.59	-0.59
casbene biosynthesis	1.4	32	3.2	582	183	582	582	0.00	0.000
fatty acid beta-oxidation II (peroxisome)	341.3	325	190.2	587	3.1	578	2935	0.94	0.674
monoterpene biosynthesis	17.5	24	39.2	575	14	0.0	4607	0.99	0.124
all-trans-farnesol biosynthesis	44.7	29	75.1	558	7	2.2	2232	0.78	0.780
geranyl diphosphate biosynthesis	114	50	68.4	554	8.1	553.7	553	0.00	0.000
palmitate biosynthesis II (bacteria and plants)	94.2	203	97.1	420	4.3	84.9	12186	0.65	0.398
jasmonic acid biosynthesis	190	124	132.6	389	2.9	361.3	7395	0.52	0.329
stearate biosynthesis II (bacteria and plants)	98.3	207	113.3	381	3.4	132.2	1904	0.34	0.212
pentose phosphate pathway (oxidative branch)	92.6	66	43.1	306	7	97.4	1224	0.59	0.591
flavonoid biosynthesis (in equisetum)	191	62	65.9	298	4.5	49.9	3572	0.007	0.001

7.3.2.2 Comparison between GC-MS result and weighted pathway approach result

For our spearmint sample, by GC-MS analysis [Figure 7.7], it is clear that the major secondary metabolites are carvone and limonene. Limonene is the precursor of carvone.

Figure 7.7 GC-MS result for spearmint sample



At the same time, in addition to carvone, there are also slightly weaker peaks related to the sesquiterpene. From our weighted pathway approach [Table 7.3], we can see the most abundance pathway in trichome is also carvone biosynthesis. In addition, the sesquiterpene-related pathways, such as all-trans-farnesol biosynthesis and trans, trans-farnesyl diphosphate biosynthesis are also present in our results, which is consistent with GC-MS analysis. However, by our weighted pathway approach, we see that diterpene, such as casbene biosynthesis, is also enriched in trichome tissue, which is not found in the GC-MS analysis. This may be because that GC-MS can only detect volatile metabolites. For non-volatile metabolites, like casbene, special gasification technique may be needed, if we use GC-MS technique.

7.3.2.3 Comparison with other pathway analysis methods

In section 6.2, we have mentioned that there are several limitations for traditional FCS methods and network-based methods. First, they do not consider dependence between each pathway. However, in our study, we use the relative expression level, not absolute expression level. Second, we have shown that pathways in our database tend to be shorter and have missing enzymes [Figure 6.5 and 6.6]. Hence, for FCS-like methods, some statistical correlation score is not applicable for our dataset. For network-based methods, they are much more applicable for longer pathways.

Therefore, we consider using mean, median, and sum value to compute overall expression level for pathways individually. For sum value, it always gives priority for longer pathways. This is because the longer a pathway is, the more enzymes are in the pathway. Mean value can be affected by some outliers, especially with only several highly expressed enzymes. Although median is less affected by outliers in large pathways, it does not take into account the precise value of each enzyme and it may not be robust for pathways that are very short.

It is very difficult to compare the results to determine which method is better, because it is difficult to say the enriched pathway set from which method is more interesting to the investigator. Commonly, most people explore the meaning of their enriched pathways just by literature search.

In our study, based on published results in mint, and in comparison with GC-MS results, transcriptome results, mean and median value for the overall expression level always produced relatively better result than sum value. When we go through the top enriched list [Table 7.3], the enriched pathways by mean and median are more reasonable for the plants we study.

7.3.2.4 Comparison between results based on absolute expression level and relative expression level

We mentioned earlier that it is reasonable to use the relative expression level, not the absolute expression level, for every enzyme in the pathway database. In order to evaluate the merit of this method, we list the top 20 enriched pathways for mint using the absolute expression level for every enzyme in Table 7.4. By comparing this result with our top 20 enriched pathways based on relative expression level in Table 7.3, we see that the major different pathways are: methionine degradation I (to homocysteine) (rank drops from 5 based on absolute expression level to 37 based on relative expression level), isoflavonoid biosynthesis I (rank drops from 13 based on absolute expression level to 403 based on relative expression level).

For methionine, the reason for not being in the top 20 pathways based on the relative expression level is because of the existence of hub compounds and hub enzymes. By the relative expression level, the overall expression level for this pathway is decreased in our results. This pathway is not related to known major secondary metabolisms for mint. Hence, it is much more biologically reasonable that this pathway is not enriched in our mint sample.

For the isoflavonoid biosynthesis I pathway, it shares one hub enzyme with flavonoid biosynthesis. By relative expression level, flavonoid biosynthesis has higher relative expression level than isoflavonoid biosynthesis for this hub enzyme. Hence, flavonoid biosynthesis is in the top 20 list based on relative expression level, but not isoflavonoid biosynthesis. Phenylpropanoids are the main mediators of plant responses to abiotic and biotic stress and they are vital to plants resistance towards pests [56], which is enriched in top 20 pathways for both results. Furthermore, phenylpropanoids serve as a rich source of metabolites for production of many other compounds like flavonoids, lignans and coumarins. Therefore, flavonoid biosynthesis, producing flavonoids, is much more reasonable than isoflavonoid biosynthesis, whose product is not related to known major secondary metabolites for mint.

Table 7.4 Top 20 enriched pathway for mint by absolute expression level for each enzyme. Trichome denotes the overall expression level using the absolute value; our method denotes overall expression level for trichome tissue based on our solution, rank is the rank for each pathway in our solution; hub compound and hub enzyme is the number for hub compound and enzyme.

pathway	trichome	trichome[our method]	rank[our method]	hub compound	hub enzyme
(4R)-carvone biosynthesis	5985.65	5864.77	1	1	1
S-adenosyl-L-methionine biosynthesis	3545.07	3545.07	2	0	0
geranyl diphosphate biosynthesis	1660.99	553.66	14	0	0
adenine and adenosine salvage VI	1648.53	1648.5	4	0	1
methionine degradation I (to homocysteine)	**1617.82**	**617.82**	**37**	**2**	**2**
menthol biosynthesis	1532.1	961.5	6	1	2
methylerythritol phosphate pathway	1483.57	1687	3	1	0
geranylgeranyldiphosphate biosynthesis	1396.59	1394.5	5	0	1
2'-deoxymugineic acid phytosiderophore biosynthesis	886.32	886.3	7	2	2
pinobanksin biosynthesis	866.62	683.3	10	2	0
pentose phosphate pathway (oxidative branch)	844.04	306.4	19	0	1
trans, trans-farnesyl diphosphate biosynthesis	744.27	744.27	9	3	2
isoflavonoid biosynthesis I	**684.1**	**0.79**	**403**	**2**	**3**
fatty acid beta-oxidation II (peroxisome)	589.88	587.124	12	3	3
casbene biosynthesis	582.01	582	11	0	0
monoterpene biosynthesis	575.88	575.83	13	0	0
free phenylpropanoid acid biosynthesis	567.92	851.885	8	1	2
all-trans-farnesol biosynthesis	558.2	558.2	14	3	3
ferulate and sinapate biosynthesis	**425.94**	**94.65**	**149**	**2**	**1**

7.3.2.5 Comparison between results based on transcriptome analysis and weighted pathway approach

By transcriptome analysis, from the top differential expression gene list, we obtain some interesting candidates, which have potentially important functions related to secondary metabolisms. By our weighted pathway approach, we can predict some new functions for our plant samples. Here, we provide some simple explanation for our weighted pathway approach in sections below.

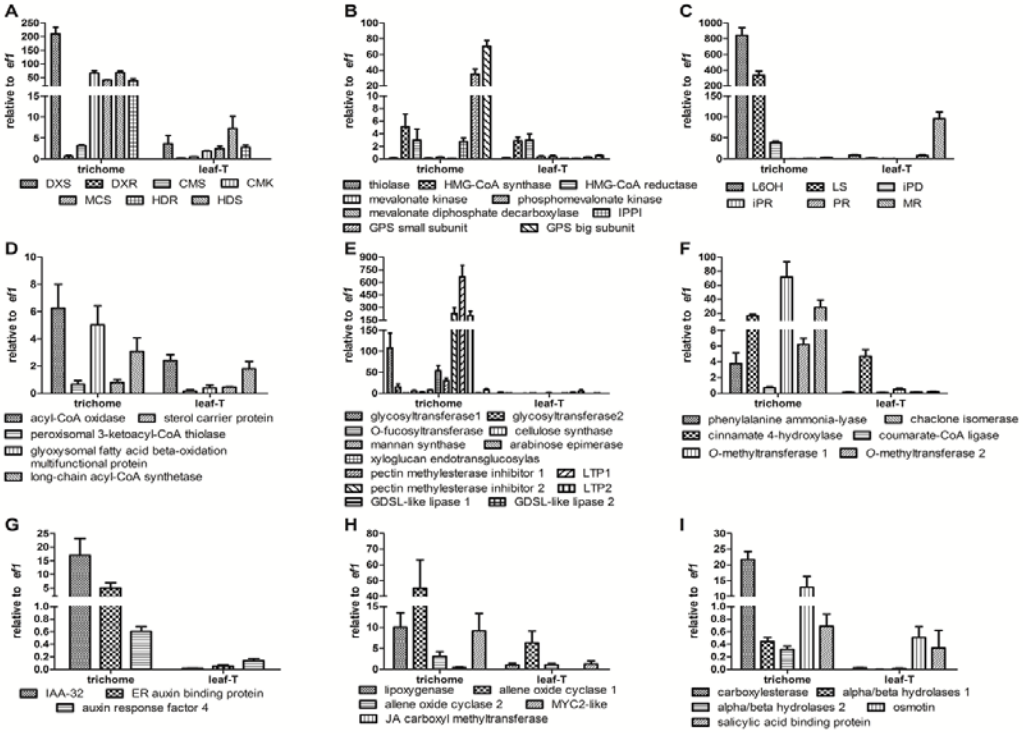
- **MEP pathway is more enriched than MVA in trichome**

After applying our weighted pathway model to our spearmint dataset, we find a lot of enriched pathways for trichome tissue, in addition to the known ones. From the result shown in Table 7.3, we can see that, not only carvone biosynthesis pathway, monoterpene, methylerythritol phosphate pathway, but also sesquiterpene (trans, trans-farnesyl diphosphate biosynthesis), diterpene (geranylgeranyldiphosphate biosynthesis), casbene biosynthesis are enriched in the trichome tissue for spearmint sample. Most of them are consistent with the metabolite results obtained from the GC-MS analysis.

From the known biosynthesis pathways shown in Figure 6.1, all isoprenoid compounds are produced by two universal 5-carbon precursors; isopentenyl diphosphate (IPP) or dimethylallyl diphosphate (DMAPP). Through evolution, two non-related biosynthetic routes have been selected for the synthesis of these two basic building blocks which use different precursors, MEP and MVA.

Based on our results, we conclude that the MEP pathway is the predominant route for spearmint. In contrast, MVA pathway is not as highly enriched as the MEP pathway. In other words, the carbon source for terpene synthesis in spearmint mainly comes from CO2 and acetyl-CoA, which is consistent with our q-PCR verification [Figure 7.8].

Figure 7.8 Q-PCR verification for several enrichment pathway predicted by our model



- **Energy production model**

Secretory trichomes are biosynthetically very active; hence, the energy requirement in these cells would presumably be more compared to other cell types. Like the result in section 7.3.1, trichome tissue does not express photosynthesis-related genes/enzymes. In addition, mint PGT trichome lacks chloroplasts. It is unknown where the energy for this specific trichome tissue comes from.

Analysis of our transcriptome data shows that most of the primary energy-producing pathways are highly enriched in PGT, like glycolysis and TCA cycle. Also, Fatty acid beta-oxidation II (peroxisome) degrading fatty acids is also enriched in PGT. Fatty acid β oxidation pathway is a process by which fatty acids are broken down to produce acetyl–coenzyme A (CoA) and it can feed the TCA cycle. Acyl-CoA oxidases (ACX) (in peroxisomes) catalyzes the first step in fatty acid β-oxidation and 3-ketoacyl-CoA thiolase (KAT) catalyzes the key step in fatty acid beta-oxidation. Gene transcripts encoding of these enzymes have been verified by q-PCR and found to be enriched in PGT [Figure 7.8]. Hence, these results show that mint trichome can obtain the necessary energy by degrading fatty acid into the TCA cycle. That also explains the high enrichment of the lipid transfer protein in trichome tissue, which may have potential function in transferring lipids from leaf to trichome.

Another evidence is that transcripts for several ABC transporters are enriched in trichome tissue, which may imply that trichome depends on the underlying leaf tissues for importing of carbon source.

In summary, for the trichome tissue in spearmint, it may have two main energy sources: carbon source from leaf tissue and degradation of fatty acid to carbon in trichomes themselves. Further experiment is needed to verify which the main source is.

- **Trichome as plants chemical defense organs**

Most glandular trichomes produce, store and secrete large amounts of different classes of secondary metabolites. The main classes of secondary chemicals that have been found to be produced in trichomes include terpenoids, phenylpropenes, methyl ketones acyl sugars and defensive proteins. All of these compounds play an important role in plant defense. Apart from having an enriched monoterpene biosynthetic pathway, our weighted pathway approach also shows the enrichment of a few other pathways (free phenylpropanoid acid biosynthesis and flavonoid biosynthesis), which are important for plant defense in spearmint PGT trichome [Table 7.3].

Phenylpropanoids are the main mediators of plant responses to abiotic and biotic stress. They are key to plant resistance towards pests [210]. Furthermore, they serve as a rich source of metabolites for production of many other compounds like lignans and coumarins. In Table 7.3, the PGT trichome shows enrichment of free phenylpropanoid acid biosynthesis, and flavonoid biosynthesis pathways. The presence of a variety of small molecular weight phenylpropanoids---e.g., caffeic, rosmarinic and ferulic acids---has been detected in leaves of different mint germplasm. Spearmint and peppermint leaves are known to produce rosmarinic acid which is a potent antioxidant [211]. Further staining for phenylpropanoids and GC-MS experiments confirm most of their presence in PGT trichome [Figure 7.7].

Additionally transcripts coding for proteinase inhibitors (PI) and polyphenols are also found to be more represented in PGTs; these are involved with defense response of plants against herbivores and pathogens.

In summary, except the known monoterpene function for trichome of spearmint, it also has other important functions, like defense tissue for plants.

7.4 Conclusion

In this chapter, our proposed weighted pathway approach is applied to the spearmint RNA-seq dataset. Comparing results obtained from GC-MS, transcriptome analysis with our weighted pathway approach, we uncover and verify several new interesting functions for the trichome tissue, such as the energy production and defense function.

We believe our weighted pathway approach will not only predict new functions/metabolites, but also provide more clues/ideas for future research about secondary metabolites.

Chapter 8

CONCLUSION

8.1 Summary

Next-generation sequencing techniques have been successfully applied in the plant metabolism community [27]. Benefitting from whole-genome sequencing techniques, after the release of the Pisifera oil palm genome, a key shell gene was found to be related to oil palm fruit formation [114]. Using RNA-seq technique, gene expression for a lot of plants, which have no reference genome yet, can be studied enabling pathway manipulation by transgenic methods. This is because there is no pre-designed probe or reference genome requirement for RNA-seq, which is different from array-based methods.

Although next-generation sequencing techniques are valuable in plant metabolism research, there are still several limitations, especially on lipid and secondary metabolisms. As the highest oil-yielding crop in the world, genome resources for oil palm are still very limited. It will be interesting to assemble genome sequence of other oil palm variants and related trees, using the released genome of Pisifera oil palm. For secondary metabolisms, using RNA-seq technique, most previous research just focus on gene level or known secondary metabolism pathways. It is important to predict new functions/metabolites for the studied plants.

We have proposed a much more comprehensive reference-based genome assembly pipeline, which is used to assemble the Dura oil palm genome. In this method, we have developed some solutions for mis-assembled scaffold and repeat scaffold identification. From the validation on a gold-standard dataset, it is clear that our pipeline outperforms DBG-based de novo assembly methods and other reference-based assembly methods.

We have generated whole-genome sequencing data for Dura oil palm and applied our reference-based genome assembly pipeline to construct a draft genome for it. This is the second sequenced genome for the oil palm community. Evaluation by three independent methods---EST coverage, genome completeness and linkage map---has demonstrated the accuracy and completeness of our draft Dura genome. We have generated RNA-seq data of 24 samples from different oil palm tissues [mesocarp, kernel, leaf, root, pollen, and flower] and developmental stages, which are helpful in the gene annotation of the draft Dura genome. Finally, around 30,000 protein-coding genes have been identified in the draft Dura genome, which is similar in size to the genome of rice [118], date palm [119] and other plants [2]. At the same time, ncRNA annotation, including tRNA, rRNA, miRNA and long noncoding RNA, are also conducted for this draft genome. Around 200 miRNA families, half of them have been verified by small RNA sequencing results, and 1,000 long noncoding RNA have been identified. In addition, by resequencing 12 different oil palm strains from three different oil palm groups: Dura, Pisifera and Tenera, we have obtained around 12 million high-quality single-nucleotide polymorphisms (SNPs). Using these population SNP data, we have identified hundreds of gene lost and appearance of start/stop codons during evolution, and thousands of genes have higher diversity sites between different oil palm groups. Some of these variants are associated with important biological features, whereas others have yet to be functionally characterized.

We have constructed an online GBrowse-based database and blast tool, which are useful for visualizing and searching genome information for oil palm. Using the database, researchers can easily visualize location information for genes, noncoding RNAs and their structures. At the same time, detail information, such as sequence, expression levels in different tissues and copy number of small RNA reads, can be visualized clearly. Using the BLAST tool, investigators can easily find homologs in oil palm, which can facilitate their experimental design and verify their hypothesis or ideas.

We have proposed a weighted pathway approach, which considers the dependency between different pathways. Finally, the relative expression level, not absolute expression level, is used to compare different pathways and samples. By validation on two different datasets, our approach is shown to be more reasonable.

We have applied this weighted pathway approach to our spearmint RNA-seq dataset, and identified several new pathways/metabolites for spearmint. At the same time, results obtained from GC-MS and Q-PCR are consistent with our prediction.

8.2 Future work

We have proposed a much more comprehensive reference-based assembly pipeline, which can utilize the genome from closely related species and reduce the depth of genome sequencing. We hope this method can help the assembly of individuals for other genetically-related species. It will be interesting to explore the genetic variation or disease variation between different individuals.

We have constructed a draft Dura genome for oil palm. Next, it will be important to identify key genes/TFs related to oil yield or oil quality. In addition, it is known that after Dura was cross pollinated with Pisifera, there was a quantum leap in oil-to-bunch from 16% (Dura) to 26% (Tenera). However, the mechanism is still unknown at the molecular level. Therefore, it is important to explore the mechanism/reason for this dramatically improvement in oil yield.

Using the identified SNPs, it is possible to select important markers for oil palm breeding. During the past thirty years, modern breeding methods based on quantitative genetics theory have been extremely successful in improving oil productivity. Hence, we hope more important markers can be identified to guide future breeding of oil palm.

For the weighted pathway approach, more plants can be used to test this approach. At the same time, it is important to perform more validation on different datasets. We hope our model can help to predict additional new functions and metabolites for different plants.

BIBLIOGRAPHY

1. Ashihara H, Crozier A, Komamine A: *Plant metabolism and biotechnology.* Cambridge ; New York: Wiley; 2011.
2. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463:**178-183.
3. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al: **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nat Biotechnol* 2012, **30:**105-111.
4. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, et al: **Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.** *Nat Genet* 2010, **42:**1053-1059.
5. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH: **Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis.** *Plant Cell* 2012, **24:**4333-4345.
6. Harada E, Kim JA, Meyer AJ, Hell R, Clemens S, Choi YE: **Expression profiling of tobacco leaf trichomes identifies genes for biotic and abiotic stresses.** *Plant Cell Physiol* 2010, **51:**1627-1637.
7. Cui H, Zhang ST, Yang HJ, Ji H, Wang XJ: **Gene expression profile analysis of tobacco leaf trichomes.** *BMC Plant Biol* 2011, **11:**76.
8. Jako C, Kumar A, Wei Y, Zou J, Barton DL, Giblin EM, Covello PS, Taylor DC: **Seed-specific over-expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight.** *Plant Physiol* 2001, **126:**861-874.
9. Wang HW, Zhang B, Hao YJ, Huang J, Tian AG, Liao Y, Zhang JS, Chen SY: **The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic Arabidopsis plants.** *Plant J* 2007, **52:**716-729.
10. Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, et al: **Sequence analysis of the genome of an oil-bearing tree, Jatropha curcas L.** *DNA Res* 2011, **18:**65-76.
11. Bouvier F, Rahier A, Camara B: **Biogenesis, molecular regulation and function of plant isoprenoids.** *Prog Lipid Res* 2005, **44:**357-429.
12. van Der Hoeven RS, Monforte AJ, Breeden D, Tanksley SD, Steffens JC: **Genetic control and evolution of sesquiterpene biosynthesis in Lycopersicon esculentum and L. hirsutum.** *Plant Cell* 2000, **12:**2283-2294.
13. Portnoy V, Benyamini Y, Bar E, Harel-Beja R, Gepstein S, Giovannoni JJ, Schaffer AA, Burger J, Tadmor Y, Lewinsohn E, Katzir N: **The molecular and biochemical basis for varietal variation in sesquiterpene content in melon (Cucumis melo L.) rinds.** *Plant Mol Biol* 2008, **66:**647-661.
14. Xia Z, Xu H, Zhai J, Li D, Luo H, He C, Huang X: **RNA-Seq analysis and de novo transcriptome assembly of Hevea brasiliensis.** *Plant Mol Biol* 2011, **77:**299-308.
15. Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, et al: **RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome.** *BMC Plant Biol* 2010, **10:**160.
16. Hu Q, Boland W, Liu JK: **6-Substituted indanoyl isoleucine conjugate induces tobacco plant responses in secondary metabolites.** *Z Naturforsch C* 2005, **60:**1-4.

17. Slocombe SP, Schauvinhold I, McQuinn RP, Besser K, Welsby NA, Harper A, Aziz N, Li Y, Larson TR, Giovannoni J, et al: **Transcriptomic and reverse genetic analyses of branched-chain fatty acid and acyl sugar production in Solanum pennellii and Nicotiana benthamiana.** *Plant Physiol* 2008, **148:**1830-1846.
18. Singh R, Ong-Abdullah M, Low ET, Manaf MA, Rosli R, Nookiah R, Ooi LC, Ooi SE, Chan KL, Halim MA, et al: **Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds.** *Nature* 2013, **500:**335-339.
19. Zheng Q, Wang XJ: **GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.** *Nucleic Acids Res* 2008, **36:**W358-363.
20. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20:**3710-3715.
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102:**15545-15550.
22. Kim SY, Volsky DJ: **PAGE: parametric analysis of gene set enrichment.** *BMC Bioinformatics* 2005, **6:**144.
23. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information.** *Bioinformatics* 2005, **21:**2988-2993.
24. Lim K, Wong L: **Finding consistent disease subnetworks using PFSNet.** *Bioinformatics* 2014, **30:**189-196.
25. Soh D, Dong D, Guo Y, Wong L: **Finding consistent disease subnetworks across microarray datasets.** *BMC Bioinformatics* 2011, **12 Suppl 13:**S15.
26. Geistlinger L, Csaba G, Kuffner R, Mulder N, Zimmer R: **From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems.** *Bioinformatics* 2011, **27:**i366-373.
27. Egan AN, Schlueter J, Spooner DM: **Applications of next-generation sequencing in plant biology.** *Am J Bot* 2012, **99:**175-185.
28. Edwards D, Batley J: **Plant genome sequencing: applications for crop improvement.** *Plant Biotechnol J* 2010, **8:**2-9.
29. Schneider GF, Dekker C: **DNA sequencing with nanopores.** *Nat Biotechnol* 2012, **30:**326-328.
30. Llaca V: **Sequencing Technologies and Their Use in Plant Biotechnology and Breeding, DNA Sequencing - Methods and Applications.** *InTech* 2012.
31. Arabidopsis Genome I: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.
32. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al: **A draft sequence of the rice genome (Oryza sativa L. ssp. indica).** *Science* 2002, **296:**79-92.
33. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al: **A draft sequence of the rice genome (Oryza sativa L. ssp. japonica).** *Science* 2002, **296:**92-100.
34. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science* 2006, **313:**1596-1604.
35. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449:**463-467.
36. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al: **The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus).** *Nature* 2008, **452:**991-996.

37. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457:**551-556.
38. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al: **The genome of the domesticated apple (Malus x domestica Borkh.).** *Nat Genet* 2010, **42:**833-839.
39. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al: **The genome of the cucumber, Cucumis sativus L.** *Nat Genet* 2009, **41:**1275-1281.
40. Wang F, Li L, Liu L, Li H, Zhang Y, Yao Y, Ni Z, Gao J: **High-throughput sequencing discovery of conserved and novel microRNAs in Chinese cabbage (Brassica rapa L. ssp. pekinensis).** *Mol Genet Genomics* 2012, **287:**555-563.
41. Potato Genome Sequencing C, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, et al: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475:**189-195.
42. Xu Q, Chen LL, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao WB, Hao BH, Lyon MP, et al: **The draft genome of sweet orange (Citrus sinensis).** *Nat Genet* 2013, **45:**59-66.
43. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al: **The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions.** *Nat Genet* 2013, **45:**51-58.
44. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al: **A whole-genome assembly of Drosophila.** *Science* 2000, **287:**2196-2204.
45. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12:**177-189.
46. Huang X, Wang J, Aluru S, Yang SP, Hillier L: **PCAP: a whole-genome assembly program.** *Genome Res* 2003, **13:**2164-2170.
47. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
48. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821-829.
49. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: de novo assembly of whole-genome shotgun microreads.** *Genome Res* 2008, **18:**810-820.
50. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19:**1117-1123.
51. Conway T, Wazny J, Bromage A, Zobel J, Beresford-Smith B: **Gossamer--a resource-efficient de novo assembler.** *Bioinformatics* 2012, **28:**1937-1938.
52. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels.** *Bioinformatics* 2012, **28:**1086-1092.
53. Ye C, Ma ZS, Cannon CH, Pop M, Yu DW: **Exploiting sparseness in de novo genome assembly.** *BMC Bioinformatics* 2012, **13 Suppl 6:**S1.
54. Peng Y, Leung HC, Yiu SM, Chin FY: **Meta-IDBA: a de Novo assembler for metagenomic data.** *Bioinformatics* 2011, **27:**i94-101.
55. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.** *Gigascience* 2012, **1:**18.
56. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M: **ABACAS: algorithm-based automatic contiguation of assembled sequences.** *Bioinformatics* 2009, **25:**1968-1969.
57. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD: **A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs.** *Nat Protoc* 2012, **7:**1260-1284.

58. Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge RL, Auvil L, Capitanu B, Zhang G, Lewin HA, Ma J: **Reference-assisted chromosome assembly.** *Proc Natl Acad Sci U S A* 2013, **110:**1785-1790.

59. Vezzi F, Cattonaro F, Policriti A: **e-RGA: enhanced Reference Guided Assembly of Complex Genomes.** *EMBnet journal* 2011, **17:**46-54.

60. Brownstein Z, Friedman LM, Shahin H, Oron-Karni V, Kol N, Abu Rayyan A, Parzefall T, Lev D, Shalev S, Frydman M, et al: **Targeted genomic capture and massively parallel sequencing to identify genes for hereditary hearing loss in Middle Eastern families.** *Genome Biol* 2011, **12:**R89.

61. Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, et al: **Whole-genome resequencing reveals loci under selection during chicken domestication.** *Nature* 2010, **464:**587-591.

62. Bick D, Dimmock D: **Whole exome and whole genome sequencing.** *Curr Opin Pediatr* 2011, **23:**594-600.

63. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, et al: **A first-generation haplotype map of maize.** *Science* 2009, **326:**1115-1117.

64. Deschamps S, la Rota M, Ratashak JP, Biddle P, Thureen D, Farmer A, Luck S, Beatty M, Nagasawa N, Michael L, et al: **Rapid Genome-wide Single Nucleotide Polymorphism Discovery in Soybean and Rice via Deep Resequencing of Reduced Representation Libraries with the Illumina Genome Analyzer.** *Plant Gen* 2010, **3:**53-68.

65. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs.** *Nat Genet* 2012, **44:**226-232.

66. Leggett RM, MacLean D: **Reference-free SNP detection: dealing with the data deluge.** *BMC Genomics* 2014, **15 Suppl 4:**S10.

67. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P: **Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences.** *Am J Bot* 2012, **99:**193-208.

68. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.

69. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29:**644-652.

70. Walter MH, Hans J, Strack D: **Two distantly related genes encoding 1-deoxy-d-xylulose 5-phosphate synthases: differential regulation in shoots and apocarotenoid-accumulating mycorrhizal roots.** *Plant J* 2002, **31:**243-254.

71. Kugler KG, Mueller LA, Graber A, Dehmer M: **Integrative network biology: graph prototyping for co-expression cancer networks.** *PLoS One* 2011, **6:**e22843.

72. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R: **A novel signaling pathway impact analysis.** *Bioinformatics* 2009, **25:**75-82.

73. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci USA* 2005, **102:**13544-13549.

74. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics* 2005, **21:**1943-1949.

75. Zampieri M, Legname G, Segre D, Altafini C: **A system-level approach for deciphering the transcriptional response to prion infection.** *Bioinformatics* 2011, **27:**3407-3414.

76. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294:**853-858.

77. Valoczi A, Hornyik C, Varga N, Burgyan J, Kauppinen S, Havelda Z: **Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide**

probes. *Nucleic Acids Res* 2004, **32:**e175.

78. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN: **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 2003, **425:**415-419.

79. Rosa A, Brivanlou AH: **microRNAs in early vertebrate development.** *Cell Cycle* 2009, **8**.

80. Liu CG, Calin GA, Volinia S, Croce CM: **MicroRNA expression profiling using microarrays.** *Nat Protoc* 2008, **3:**563-578.

81. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.

82. Wang H, Zhang X, Liu J, Kiba T, Woo J, Ojo T, Hafner M, Tuschl T, Chua NH, Wang XJ: **Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions.** *Plant J* 2011, **67:**292-304.

83. An J, Lai J, Lehman ML, Nelson CC: **miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data.** *Nucleic Acids Res* 2013, **41:**727-737.

84. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, Schwach F, Dalmay T, Moulton V: **The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets.** *Bioinformatics* 2012, **28:**2059-2061.

85. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458:**223-227.

86. Bentley DR: **Whole-genome re-sequencing.** *Curr Opin Genet Dev* 2006, **16:**545-552.

87. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, et al: **Single-molecule DNA sequencing of a viral genome.** *Science* 2008, **320:**106-109.

88. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: **De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18:**802-809.

89. Simpson JT, Durbin R: **Efficient de novo assembly of large genomes using compressed data structures.** *Genome Res* 2012, **22:**549-556.

90. Narzisi G, Mishra B: **Comparing de novo genome assembly: the long and short of it.** *PLoS One* 2011, **6:**e19175.

91. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85:**2444-2448.

92. Rasmussen KR, Stoye J, Myers EW: **Efficient q-gram filters for finding all epsilon-matches over a given length.** *J Comput Biol* 2006, **13:**296-308.

93. Idury RM, Waterman MS: **A new algorithm for DNA sequence assembly.** *J Comput Biol* 1995, **2:**291-306.

94. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci USA* 2001, **98:**9748-9753.

95. Chikhi R, Rizk G: **Space-efficient and exact de Bruijn graph representation based on a Bloom filter.** *Algorithms Mol Biol* 2013, **8:**22.

96. Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K, et al: **The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla).** *Nat Genet* 2013, **45:**456-461, 461e451-452.

97. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463:**311-317.

98. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24:**133-141.

99. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, et al: **Reference-guided assembly of four diverse Arabidopsis thaliana genomes.** *Proc Natl Acad Sci USA* 2011, **108:**10249-10254.

100. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng HW: **Comparative studies of de novo**

assembly tools for next-generation sequencing technologies. *Bioinformatics* 2011, **27:**2031-2037.

101. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.

102. Soderlund C, Bomhoff M, Nelson WM: **SyMAP v3.4: a turnkey synteny system with application to plant genomes.** *Nucleic Acids Res* 2011, **39:**e68.

103. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, et al: **Assemblathon 1: a competitive assessment of de novo short read assembly methods.** *Genome Res* 2011, **21:**2224-2241.

104. Lee KR, Kim SH, Go YS, Jung SM, Roh KH, Kim JB, Suh MC, Lee S, Kim HU: **Molecular cloning and functional analysis of two FAD2 genes from American grape (Vitis labrusca L.).** *Gene* 2012, **509:**189-194.

105. Pham AT, Shannon JG, Bilyeu KD: **Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil.** *Theor Appl Genet* 2012, **125:**503-515.

106. Wang ML, Barkley NA, Chen Z, Pittman RN: **FAD2 gene mutations significantly alter fatty acid profiles in cultivated peanuts (Arachis hypogaea).** *Biochem Genet* 2011, **49:**748-759.

107. Cao S, Zhou XR, Wood CC, Green AG, Singh SP, Liu L, Liu Q: **A large and functionally diverse family of Fad2 genes in safflower (Carthamus tinctorius L.).** *BMC Plant Biol* 2013, **13:**5.

108. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112-1115.

109. Zhang H, Miao H, Wang L, Qu L, Liu H, Wang Q, Yue M: **Genome sequencing of the important oilseed crop Sesamum indicum L.** *Genome Biol* 2013, **14:**401.

110. Huang YY, Matzke AJ, Matzke M: **Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (Cocos nucifera).** *PLoS One* 2013, **8:**e74736.

111. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences.** *Plant J* 2008, **53:**661-673.

112. Ruuska SA, Schwender J, Ohlrogge JB: **The capacity of green oilseeds to utilize photosynthesis to drive biosynthetic processes.** *Plant Physiol* 2004, **136:**2700-2709.

113. Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoocha T, Jomchai N, Jantasuriyarat C, Tragoonrung S, Tangphatsornruang S: **Characterization of the chloroplast genome sequence of oil palm (Elaeis guineensis Jacq.).** *Gene* 2012, **500:**172-180.

114. Singh R, Low ET, Ooi LC, Ong-Abdullah M, Ting NC, Nagappan J, Nookiah R, Amiruddin MD, Rosli R, Manaf MA, et al: **The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK.** *Nature* 2013, **500:**340-344.

115. Gao S, Sung WK, Nagarajan N: **Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences.** *J Comput Biol* 2011, **18:**1681-1691.

116. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27:**578-579.

117. Dayarian A, Michael TP, Sengupta AM: **SOPRA: Scaffolding algorithm for paired reads via statistical optimization.** *BMC Bioinformatics* 2010, **11:**345.

118. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al: **A draft sequence of the rice genome (Oryza sativa L. ssp. indica).** *Science* 2002, **296:**79-92.

119. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, et al: **De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera).** *Nat Biotechnol* 2011, **29:**521-527.

120. Bourgis F, Kilaru A, Cao X, Ngando-Ebongue GF, Drira N, Ohlrogge JB, Arondel V: **Comparative transcriptome and metabolite analysis of oil palm and date palm**

mesocarp that differ dramatically in carbon partitioning. *Proc Natl Acad Sci USA* 2011, **108:**12527-12532.

121. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.

122. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23:**1061-1067.

123. Seng TY, Mohamed Saad SH, Chin CW, Ting NC, Harminder Singh RS, Qamaruz Zaman F, Tan SG, Syed Alwee SS: **Genetic linkage map of a high yielding FELDA delixyangambi oil palm cross.** *PLoS One* 2011, **6:**e26593.

124. Ting NC, Zaki NM, Rosli R, Low ET, Ithnin M, Cheah SC, Tan SG, Singh R: **SSR mining in oil palm EST database: application in oil palm germplasm diversity studies.** *J Genet* 2010, **89:**135-145.

125. Billotte N, Marseillac N, Risterucci AM, Adon B, Brottier P, Baurens FC, Singh R, Herran A, Asmady H, Billot C, et al: **Microsatellite-based high density linkage map in oil palm (Elaeis guineensis Jacq.).** *Theor Appl Genet* 2005, **110:**754-765.

126. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.

127. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21 Suppl 1:**i351-358.

128. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35:**W265-268.

129. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.

130. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.

131. **Genome sequencing and analysis of the model grass Brachypodium distachyon.** *Nature* 2010, **463:**763-768.

132. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19 Suppl 2:**ii215-225.

133. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

134. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training.** *Genome Res* 2008, **18:**1979-1990.

135. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2012, **40:**D1202-1210.

136. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.

137. Jin J, Liu J, Wang H, Wong L, Chua NH: **PLncDB: plant long non-coding RNA database.** *Bioinformatics* 2013, **29:**1068-1071.

138. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.

139. Conesa A, Gotz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics.** *Int J Plant Genomics* 2008, **2008:**619832.

140. Jouannic S, Argout X, Lechauve F, Fizames C, Borgel A, Morcillo F, Aberlenc-Bertossi F, Duval Y, Tregear J: **Analysis of expressed sequence tags from oil palm (Elaeis guineensis).** *FEBS Lett* 2005, **579:**2709-2714.

141. Washietl S, Will S, Hendrix DA, Goff LA, Rinn JL, Berger B, Kellis M: **Computational analysis of noncoding RNAs.** *Wiley Interdiscip Rev RNA* 2012, **3:**759-778.

142. Fichant GA, Burks C: **Identifying potential tRNA genes in genomic DNA sequences.** *J Mol Biol* 1991, **220:**659-671.

143. Pavesi A, Conterio F, Bolchi A, Dieci G, Ottonello S: **Identification of new eukaryotic**

tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res* 1994, **22:**1247-1256.

144. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.

145. Schattner P, Brooks AN, Lowe TM: **The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.** *Nucleic Acids Res* 2005, **33:**W686-689.

146. **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.

147. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A: **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Res* 2013, **41:**D226-232.

148. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25:**1335-1337.

149. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res* 2011, **39:**D152-157.

150. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA.** *Mol Cell* 2004, **14:**787-799.

151. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.

152. Schattner P, Decatur WA, Davis CA, Ares M, Jr., Fournier MJ, Lowe TM: **Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome.** *Nucleic Acids Res* 2004, **32:**4281-4296.

153. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, et al: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.** *Nature* 2010, **464:**1071-1076.

154. Hu W, Yuan B, Flygare J, Lodish HF: **Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation.** *Genes Dev* 2011, **25:**2573-2578.

155. Wen J, Parker BJ, Weiller GF: **In Silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula.** *In Silico Biol* 2007, **7:**485-505.

156. Boerner S, McGinnis KM: **Computational identification and functional predictions of long noncoding RNA in Zea mays.** *PLoS One* 2012, **7:**e43047.

157. Xin M, Wang Y, Yao Y, Song N, Hu Z, Qin D, Xie C, Peng H, Ni Z, Sun Q: **Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing.** *BMC Plant Biol* 2011, **11:**61.

158. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci USA* 2009, **106:**11667-11672.

159. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

160. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7:**562-578.

161. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

162. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, et al: **A map of rice genome variation reveals the origin of cultivated rice.** *Nature* 2012, **490:**497-501.

163. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, et al: **Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana.** *Science* 2007, **317:**338-342.

164. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using**

multilocus genotype data. *Genetics* 2000, **155:**945-959.

165. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4:**e72.

166. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.

167. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26:**2204-2207.

168. Glas JJ, Schimmel BC, Alba JM, Escobar-Bravo R, Schuurink RC, Kant MR: **Plant glandular trichomes as targets for breeding or engineering of resistance to herbivores.** *Int J Mol Sci* 2012, **13:**17077-17103.

169. Mathur J, Chua NH: **Microtubule stabilization leads to growth reorientation in Arabidopsis trichomes.** *Plant Cell* 2000, **12:**465-477.

170. Larkin JC, Brown ML, Schiefelbein J: **How do cells know what they want to be when they grow up? Lessons from epidermal patterning in Arabidopsis.** *Annu Rev Plant Biol* 2003, **54:**403-430.

171. Trethewey RN, Krotzky AJ, Willmitzer L: **Metabolic profiling: a Rosetta Stone for genomics?** *Curr Opin Plant Biol* 1999, **2:**83-85.

172. Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A: **Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems.** *Plant Cell* 2001, **13:**11-29.

173. Fiehn O: **Metabolic networks of Cucurbita maxima phloem.** *Phytochemistry* 2003, **62:**875-886.

174. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Comput Biol* 2012, **8:**e1002375.

175. Tranbarger TJ, Dussert S, Joet T, Argout X, Summo M, Champion A, Cros D, Omore A, Nouy B, Morcillo F: **Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism.** *Plant Physiol* 2011, **156:**564-584.

176. Shearman JR, Jantasuriyarat C, Sangsrakru D, Yoocha T, Vannavichit A, Tragoonrung S, Tangphatsornruang S: **Transcriptome analysis of normal and mantled developing oil palm flower and fruit.** *Genomics* 2013, **101:**306-312.

177. Feng C, Chen M, Xu CJ, Bai L, Yin XR, Li X, Allan AC, Ferguson IB, Chen KS: **Transcriptomic analysis of Chinese bayberry (Myrica rubra) fruit development and ripening using RNA-Seq.** *BMC Genomics* 2012, **13:**19.

178. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.

179. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.

180. D'Haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16:**707-726.

181. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302:**249-255.

182. Jiang Z, Gentleman R: **Extensions to gene set enrichment.** *Bioinformatics* 2007, **23:**306-313.

183. Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data.** *Stat Appl Genet Mol Biol* 2004, **3:**Article16.

184. Sivachenko AY, Yuryev A, Daraselia N, Mazo I: **Molecular networks in microarray analysis.** *J Bioinform Comput Biol* 2007, **5:**429-456.

185. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S,

Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-484.

186. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, et al: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8:**R39.

187. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2012, **40:**D742-753.

188. Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY: **MetaCyc and AraCyc. Metabolic pathway databases for plant research.** *Plant Physiol* 2005, **138:**27-37.

189. Urbanczyk-Wochniak E, Sumner LW: **MedicCyc: a biochemical pathway database for Medicago truncatula.** *Bioinformatics* 2007, **23:**1418-1423.

190. May P, Christian JO, Kempa S, Walther D: **ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii.** *BMC Genomics* 2009, **10:**209.

191. Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C, et al: **Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants.** *Plant Physiol* 2010, **153:**1479-1491.

192. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.

193. tom Dieck H, Doring F, Fuchs D, Roth HP, Daniel H: **Transcriptome and proteome analysis identifies the pathways that increase hepatic lipid accumulation in zinc-deficient rats.** *J Nutr* 2005, **135:**199-205.

194. Havaux M, Eymery F, Porfirova S, Rey P, Dormann P: **Vitamin E protects against photoinhibition and photooxidative stress in Arabidopsis thaliana.** *Plant Cell* 2005, **17:**3451-3469.

195. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4:**249-264.

196. Hollander-Czytko H, Grabowski J, Sandorf I, Weckermann K, Weiler EW: **Tocopherol content and activities of tyrosine aminotransferase and cystine lyase in Arabidopsis under stress conditions.** *J Plant Physiol* 2005, **162:**767-770.

197. Riewe D, Koohi M, Lisec J, Pfeiffer M, Lippmann R, Schmeichel J, Willmitzer L, Altmann T: **A tyrosine aminotransferase involved in tocopherol synthesis in Arabidopsis.** *Plant J* 2012, **71:**850-859.

198. Sandorf I, Hollander-Czytko H: **Jasmonate is involved in the induction of tyrosine aminotransferase and tocopherol biosynthesis in Arabidopsis thaliana.** *Planta* 2002, **216:**173-179.

199. Nawrath C, Metraux JP: **Salicylic acid induction-deficient mutants of Arabidopsis express PR-2 and PR-5 and accumulate high levels of camalexin after pathogen inoculation.** *Plant Cell* 1999, **11:**1393-1404.

200. Garcion C, Lohmann A, Lamodiere E, Catinot J, Buchala A, Doermann P, Metraux JP: **Characterization and biological function of the ISOCHORISMATE SYNTHASE2 gene of Arabidopsis.** *Plant Physiol* 2008, **147:**1279-1287.

201. Schlaeppi K, Abou-Mansour E, Buchala A, Mauch F: **Disease resistance of Arabidopsis to Phytophthora brassicae is established by the sequential action of indole glucosinolates and camalexin.** *Plant J* 2010, **62:**840-851.

202. Lange BM, Mahmoud SS, Wildung MR, Turner GW, Davis EM, Lange I, Baker RC, Boydston RA, Croteau RB: **Improving peppermint essential oil yield and composition by metabolic engineering.** *Proc Natl Acad Sci U S A* 2011, **108:**16944-16949.

203. Lange BM, Wildung MR, Stauber EJ, Sanchez C, Pouchnik D, Croteau R: **Probing essential oil biosynthesis and secretion by functional evaluation of expressed sequence tags**

from mint glandular trichomes. *Proc Natl Acad Sci U S A* 2000, **97:**2934-2939.

204. Croteau RB, Davis EM, Ringer KL, Wildung MR: **(-)-Menthol biosynthesis and molecular genetics.** *Naturwissenschaften* 2005, **92:**562-577.

205. Champagne A, Boutry M: **Proteomic snapshot of spearmint (Mentha spicata L.) leaf trichomes: a genuine terpenoid factory.** *Proteomics* 2013, **13:**3327-3332.

206. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61-65.

207. Turner GW, Gershenzon J, Croteau RB: **Distribution of peltate glandular trichomes on developing leaves of peppermint.** *Plant Physiol* 2000, **124:**655-664.

208. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.

209. Morohashi K, Grotewold E: **A systems approach reveals regulatory circuitry for Arabidopsis trichome initiation by the GL3 and GL1 selectors.** *PLoS Genet* 2009, **5:**e1000396.

210. La Camera S, Gouzerh G, Dhondt S, Hoffmann L, Fritig B, Legrand M, Heitz T: **Metabolic reprogramming in plant innate immunity: the contributions of phenylpropanoid and oxylipin pathways.** *Immunol Rev* 2004, **198:**267-284.

211. Tahira R, Naeemullah M, Akbar F, Masood MS: **Major Phenolic Acids of Local and Exotic Mint Germplasm Grown in Islamabad.** *Pakistan Journal of Botany* 2011, **43:**151-154.