<table>
<thead>
<tr>
<th>Title</th>
<th>Molecular evolution of imprinted, orphan and De novo genes in plant genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Tuteja, Reetu</td>
</tr>
<tr>
<td>Publication Date</td>
<td>2016-02-23</td>
</tr>
<tr>
<td>Item record</td>
<td><a href="http://hdl.handle.net/10379/5786">http://hdl.handle.net/10379/5786</a></td>
</tr>
</tbody>
</table>
Molecular Evolution of Imprinted, Orphan and De novo Genes in Plant Genomes

Volume 1 of 1

Reetu Tuteja, M.Sc.

A thesis submitted to the National University of Ireland, Galway

For the degree of Doctor of Philosophy

College of Science, School of Natural Sciences

Plant and AgriBiosciences Research Centre

Under the supervision of Professor Charles Spillane

Head of School of Natural Sciences – Professor Vincent O’Flaherty

September 2015
Table of Contents

Table of Contents ........................................................................................................... i
List of Publications and Conferences ........................................................................ v
List of Tables .................................................................................................................. vii
List of Figures ................................................................................................................ ix
Abbreviations ................................................................................................................ xii
Declaration .................................................................................................................... xiii
Acknowledgments ......................................................................................................... xiv
Summary ......................................................................................................................... 1

Chapter 1 . Introduction ................................................................................................. 3
  1.1 Origination of new genes ....................................................................................... 4
    1.1.1 Mechanisms of origination of new genes ...................................................... 5
  1.2 Duplication and origin of new genes ................................................................... 8
    1.2.1 Generation of duplicate genes ...................................................................... 8
        1.2.1.1 Duplication of the entire genome ......................................................... 8
        1.2.1.2 Duplication of a single chromosome .................................................. 11
        1.2.1.3 Duplication of single gene or group of genes ...................................... 11
    1.2.2 Evolutionary fate of duplicate genes .............................................................. 11
  1.3 Orthologs, paralogs and their functional conservation ........................................ 13
  1.4 Orphan genes and De novo genes ....................................................................... 17
  1.5 Fate of newly arisen mutations ........................................................................... 21
    1.5.1 Methods for identifying selection pressures acting on protein-coding
        sequences ........................................................................................................... 22
        1.5.1.1 Maximum likelihood method for measuring selection pressure ....... 23
  1.6 Dosage Effects ..................................................................................................... 27
    1.6.1 Genomic imprinting ...................................................................................... 29
        1.6.1.1 Imprinted genes in plants ................................................................... 29
        1.6.1.2 Evolution of genomic imprinting ..................................................... 31
  1.7 Aim of Thesis ....................................................................................................... 33

Chapter 2 . Orphan genes and de novo originated overprinted genes identified in
  plant species ................................................................................................................. 35
  2.1 Introduction ........................................................................................................... 36
2.2 Materials and methods .................................................................................................40
  2.2.1 Identification of orphan genes in *S. bicolor* and *C. sativa* ...............................40
  2.2.2 Identification of evolutionary origins of orphan genes .......................................40
  2.2.3 Expression evidence of sorghum-specific orphans in RNA-seq data of 40 accessions .................................................................................................................................................41
  2.2.4 Identification of overprinted genes in *A. thaliana* ............................................41
  2.2.5 Determining the gene age and rate of evolution of overprinted genes .............42

2.3 Results ..........................................................................................................................44
  2.3.1 Identification of 3,761 Brassicaceae-specific orphan genes in *C. sativa* ......44
  2.3.2 Identification of 2,499 species-specific orphan genes in *S. bicolor* ...............47
  X-axis represents the 40 accessions of *S. bicolor*, % of expressed orphan genes in each accession are shown on Y-axis ..................................................................................................................50
  2.3.3 Accelerated evolutionary rate of De novo originated overprinted genes in *A. thaliana* ..................................................................................................................................................51

2.4 Discussion .....................................................................................................................53

Chapter 3. Cytoplasmic male sterility associated chimeric ORFs identified by mitochondrial genome sequencing of four *Cajanus* genotypes ......................................................56

Abstract ..............................................................................................................................57

3.1 Introduction ....................................................................................................................58

3.2 Materials and methods ...............................................................................................60
  3.2.1 Plant material and mitochondrial DNA isolation ...............................................60
  3.2.2 Sequencing and assembly ..................................................................................60
  3.2.4 Gene order comparison ....................................................................................62
  3.2.5 Genome alignment and representation .............................................................62
  3.2.6 Identification of rearrangements and no-coverage regions ..............................63
  3.2.7 Chimeric ORFs .................................................................................................63

3.3 Results and discussion .................................................................................................64
  3.3.1 Sequencing and assembly of four mitochondrial genomes of pigeonpea...64
  3.3.2 Finishing of reference mitochondrial genome ICPA 2039 ..............................67
  3.3.3 Gene content ....................................................................................................69
  3.3.4 Structural features of the pigeonpea mitochondrial genome compared to other plant species ..........................................................................................................................70
  3.3.5 Comparison of gene order with other plant species .........................................72
3.3.6 Comparison of mitochondrial genome sequences among Cajanus lines ... 74
3.3.7 Candidate CMS-associated chimeric ORFs ........................................ 75

Chapter 4 . Construction of a plant orthology database ................................ 80
  4.1 Introduction ................................................................................................. 81
  4.2 Materials and Methods ............................................................................ 83
    4.2.1 Data ...................................................................................................... 83
    4.2.2 Construction of orthology database .................................................... 83
  4.3 Results and example uses of the database .............................................. 87
    4.3.1 Description of the orthology database .............................................. 87
    4.3.2 Using the orthology database ............................................................. 87
  4.4 Conclusions ................................................................................................. 88

Chapter 5 . Signatures of positive Darwinian selection in Arabidopsis thaliana
  paternally expressed imprinted genes ......................................................... 89
  5.1 Introduction ................................................................................................. 90
  5.2 Materials and methods ............................................................................ 93
    5.2.1 Data ...................................................................................................... 93
    5.2.2 Gene age and evolutionary rate ............................................................ 93
    5.2.3 Orthologs Identification ...................................................................... 93
    5.2.4 Random Sets ....................................................................................... 94
    5.2.5 Multiple Sequence Alignment ........................................................... 94
    5.2.6 Recombination Detection .................................................................... 94
    5.2.7 Tree building ....................................................................................... 95
    5.2.8 Selection Analysis ............................................................................... 95
    5.2.9 Population Genetic Analysis .............................................................. 96
  5.3 Results ......................................................................................................... 97
    5.3.1 iPEGs and iMEGs showed similar age distribution and evolutionary rates,
        but accelerated evolutionary rates compared to whole genome .............. 97
    5.3.2 A. thaliana imprinted genes show enrichment for recent whole genome
duplication events (alpha) ............................................................................ 99
    5.3.3 Imprinted genes are under positive selection ....................................... 101
    5.3.4 Positive selection in imprinted genes compared to random set selected from
genome and endosperm ............................................................................. 110
5.3.5 Most of the imprinted genes displayed fixation of positively selected sites (posterior probability > 0.95) within \( A. thaliana \) population .............................................. 113
5.3.6 Both iPEGs and iMEGs showed an excess of deleterious mutations in \( A. thaliana \) population data ................................................................. 117
5.3.7 Imprinted genes with disrupted reading frame in \( A. thaliana \) accessions showed high \( Dn/Ds \) and high \( Pn/Ps \) ratios .............................................. 118
5.4 Discussion .................................................................................. 120
5.5 Conclusions and future work .................................................. 122

Chapter 6. Detection of imprinting and dosage effects in triploid \( Arabidopsis thaliana \) embryos ............................................................................. 123

6.1 Introduction .............................................................................. 124
6.2 Materials and Methods ............................................................... 127

   6.2.1 High-throughput sequencing and data filtering ....................... 127
   6.2.2 SNP data .................................................................................. 127
   6.2.3 Mapping reads on Col-0 and C24 genomes ......................... 127
   6.2.4 Determining Allele-specific expression ............................... 128

6.3 Results ........................................................................................ 129

   6.3.1 A significantly higher number of genes showed maternal bias in both the maternal excess and paternal excess triploids ......................... 129
   6.3.2 Gene-specific imprinting observed in reciprocal F1 triploid hybrid embryos ................................................................. 132
   6.3.3 Overlap of candidate imprinted genes identified in F1 triploid embryos with the known imprinted genes in \( A. thaliana \) endosperm.............................. 134

6.4 Discussion .................................................................................. 135
6.5 Conclusions and future plan .................................................. 137

Chapter 7. Summary and future directions ..................................... 138

Supplementary Table and Figures .................................................. 143

Bibliography .................................................................................... 198
List of Publications and Conferences

Publications


Posters/Presentations

- 8-10 October 2011: Poster presentation ‘Comparative analysis of four mitochondrial genomes to understand the molecular basis of CMS in pigeonpea (Cajanus spp.)’ at RECOMB Satellite Workshop on Comparative Genomics, Galway, Ireland.
3-5 April 2012: Oral Presentation ‘Comparative analysis of four mitochondrial genomes to understand the molecular basis of CMS in pigeonpea (Cajanus spp.)’ at IPSAM meeting, Dublin, Ireland.

16-17 May 2013: Oral Presentation 'Pigeonpea improvement by sequencing of nuclear and mitochondrial genomes' at IPSAM meeting, Galway, Ireland.


11-12 May 2015: Oral Presentation ‘Molecular evolution and expression analysis of orphan genes in the Sorghum bicolor genome’ at IPSAM meeting, Maynooth, Ireland.

**PhD Course work**

- Genomic Technologies (2012)

- Graduate Research Information Skills (2013)

- Journal Club (2013)

- Participation in Workshops/Courses (2013)


- Oral/Poster Communications (2014)

**Workshop Attended**

- Statistics and Computing in Genome Data Science, CSAMA, Brixen, Italy, 1-6 July 2012
# List of Tables

Table 1.1: Calculation for Likelihood Ratio test (LRT) .................................................................25  
Table 2.1: Comparison of sequence features of orphans and non-orphans in the reference genome of *C. sativa* ........................................................................................................................................44  
Table 2.2: Evolutionary origins of orphan genes in *C. sativa* ......................................................46  
Table 2.3: Comparison of sequence features of orphans and non-orphan genes in the reference genome of *Sorghum bicolor* ........................................................................................................48  
Table 2.4: Modes of evolutionary origin of orphan genes in *S. bicolor* ..........................................49  
Table 2.5: Evolutionary rate of *A. thaliana* overprinted genes in comparison to the whole genome as background ..........................................................................................................................51  
Table 3.1: Generation of 454/FLX data and assembly statistics of ICPA, ICPB, ICPH, ICPW ........................................................................................................................................66  
Table 3.2: Genome coverage by coding features in ICPA 2039 mitochondrial genome assembly ...........................................................................................................................................69  
Table 3.3: Potential chimeric ORFs identified from the no coverage and rearrangement regions between ICPA 2039 and ICPB 2039 line ........................................................................................................79  
Table 4.1: List of 34 sequenced plant and algal species that were used to construct plant orthology database ..................................................................................................................................84  
Table 4.2: Statistics of orthology database of 34 sequenced species ...............................................87  
Table 5.1: Evolutionary rate of imprinted genes in comparison to the whole genome as background ............................................................................................................................98  
Table 5.2: Number of iMEGs and iPEGs identified to be under lineage-specific selection and site-specific using codeML models ..........................................................................................103  
Table 5.3: Sites in imprinted genes identified to be under lineage-specific positive selection ..........................................................................................................................104  
Table 5.4: Sites in imprinted genes identified to be under site-specific positive selection ..........................................................................................................................107  
Table 5.5: Percentage of fixation of lineage-specific positively selected sites (posterior probability > 0.95) in Ath population data ..........................................................................................113  
Table 5.6: Percentage of conservation of lineage-specific positively selected sites in Ath 80 population data identified in AT1G61330 ..................................................................................115
Table 5.7: Number of non-synonymous and synonymous substitutions, Observed $FI$ and $eFI$ for $A. thaliana$ imprinted genes. ..............................................................................................117
Table 5.8: High $Dn/Ds$ and High $Pn/Ps$ of $A. thaliana$ imprinted genes identified as pseudogenes. ........................................................................................................................................118
Table 6.1: Genes identified to be imprinted (MEGs and PEGs) in maternal and paternal excess triploid data..............................................................................................................................132
Table 6.2: Overlap of candidate imprinted genes identified in triploid embryo with the known imprinted genes in $A. thaliana$ endosperm..............................................................................................134
List of Figures

Figure 1.1: Mechanisms of origin of new genes.................................................................6
Figure 1.2: Schematic representation of life cycle of genes..............................................7
Figure 1.3: Representation of the putative times of ancestral polyploidy events inferred for flowering plants.................................................................................................9
Figure 1.4: Phylogeny of sequenced plant and green algae genomes...............................10
Figure 1.5: Example tree to explain the concept of inparalogs, co-orthologs, outparalogs...............................................................................................................................14
Figure 1.6: Results of different orthology prediction methods assessed using a benchmark dataset of 100 genes from the TreeFam-A resource...........................................16
Figure 1.7: Orphan genes as part of the variable genome..................................................18
Figure 1.8: Explains a model of variable $\omega$ ratios among branches of a tree of four sequences.........................................................................................................................23
Figure 1.9: Representation of the types of effects observed for gene expression in aneuploids..........................................................................................................................28
Figure 1.10: Imprinted genes in A. thaliana identified by different high-throughput screens showed limited overlap (Source: McKeown et al. 2013).............................................31
Figure 2.1: An example of a representative overprinted gene in A. thaliana....................38
Figure 2.2: Areas of origin and development for the domesticated races of bicolor and their possible migration routes (Source: Kimber et al., 2013)..................................................39
Figure 2.3: Schematic of search strategy for identification of orphan genes in a given genome............................................................................................................................43
Figure 2.4: Percentage of orphan genes on 20 chromosomes of C. sativa.........................45
Figure 2.5: Orphan genes (as percentage of genes on each chromosome) located on each of the 10 chromosomes of sorghum ($\chi^2$-test, $p=0.9514$, df=10).........................................48
Figure 2.6: Percentage of orphan genes expressed in each of the S. bicolor accessions. ..............................................................................................................................................50
Figure 2.7: Number of overlapping gene pairs identified in the A. thaliana reference genome..............................................................................................................................52
Figure 3.1: A scheme showing linking scaffold with the help of graph in a preliminary view of the assembly...........................................................................................................61
Figure 3.2: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Vigna radiata.* ................................................................. 73
Figure 3.3: Alignments of *Cajanus* mitochondrial genomes. ................................. 77
Figure 4.1: Pipeline for identification of orthology clusters in sequenced plant species. ...................................................................................................................... 86
Figure 5.1: Distribution of age of iMEGs and iPEGs. ................................................. 98
Figure 5.2: Number of species in orthology clusters of imprinted genes. .............. 100
Figure 5.3: Schematic for identification of evolution of 148 *A. thaliana* imprinted genes. ............................................................................................................ 101
Figure 5.4: Venn diagram for imprinted genes found be both under lineage-specific selection and site-specific selection using codeML models. ...................... 102
Figure 5.5: Represents the percentage of lineage-specific positively selected genes identified in imprinted genes and random sets. ........................................ 111
Figure 5.6: Represents the percentage of site-specific positively selected genes identified in imprinted genes and random sets. ........................................... 112
Figure 5.7: Positively selected sites identified for AT1G61330.............................. 116
Figure 5.8: 2D plot of $Dn/Ds$ and $Pn/Ps$ of *A. thaliana* protein-coding genes. ...... 119
Figure 6.1: Allelic contributions of three genes in reciprocal triploid hybrids using Col-0 as a tetraploid parent (experimental work conducted by Antoine Fort). .... 126
Figure 6.2: Allelic contributions of three genes in reciprocal triploid hybrids using C24 as a tetraploid parent (experimental work conducted by Antoine Fort) .... 126
Figure 6.3: Percentage of genes showed dominant gene expression from Col-0 and C24. .................................................................................................................. 130
Figure 6.4: Percentage of genes that showed paternal or maternal bias in maternal excess triploids (4xC24 X 2xCol) and paternal excess triploids (2xC24 X 4xC24). 131
Figure 6.5: Distribution of allelic contribution of biased and unbiased genes in transcriptomes of paternal excess and maternal excess triploids. ....................... 133
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>Allele-specific expression</td>
</tr>
<tr>
<td>Ath</td>
<td><em>Arabidopsis thaliana</em></td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>BEB</td>
<td>Bayes Empirical Bayes</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>Bp</td>
<td>Base pairs</td>
</tr>
<tr>
<td>Col-0</td>
<td>Columbia (ecotype of <em>Arabidopsis thaliana</em>)</td>
</tr>
<tr>
<td>CMS</td>
<td>Cytoplasmic male sterility</td>
</tr>
<tr>
<td>Dn</td>
<td>Non-synonymous substitution per non-synonymous site</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribose nucleic acid</td>
</tr>
<tr>
<td>Ds</td>
<td>Synonymous substitution per synonymous site</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>HGT</td>
<td>Horizontal Gene Transfer</td>
</tr>
<tr>
<td>ICHEC</td>
<td>Irish Centre for High End Computing</td>
</tr>
<tr>
<td>IMEG</td>
<td>Imprinted maternally expressed gene</td>
</tr>
<tr>
<td>IPEG</td>
<td>Imprinted paternally expressed gene</td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood Ratio Tests</td>
</tr>
<tr>
<td>MCL</td>
<td>Markov Clustering Algorithm</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>mtDNA</td>
<td>Mitochondrial DNA</td>
</tr>
<tr>
<td>mtGene</td>
<td>Mitochondrial Gene</td>
</tr>
<tr>
<td>Mya</td>
<td>Million years ago</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PAML</td>
<td>Phylogenetic Analysis by Maximum Likelihood</td>
</tr>
<tr>
<td>PP</td>
<td>Posterior Probability</td>
</tr>
<tr>
<td>RAxML</td>
<td>Randomized Axelerated Maximum Likelihood</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SGO</td>
<td>Single gene orthologs</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>TAIR</td>
<td>The Arabidopsis Information Resource</td>
</tr>
<tr>
<td>TRG</td>
<td>Taxonomically restricted gene</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
<tr>
<td>WT</td>
<td>Wild type</td>
</tr>
</tbody>
</table>
Declaration

I certify that this thesis is my own work, and that I have not used this work in the course of another degree, either at National University of Ireland Galway, or elsewhere.

Signed:

___________________
Reetu Tuteja
Acknowledgments

I sincerely acknowledge my supervisor Professor Charles Spillane, for giving me an opportunity to pursue PhD in his group, for his guidance and support. I am grateful to him for continuous encouragement, optimism and above all his patience. I also thank Professor Rajeev Varshney who I worked with in ICRISAT before starting my PhD and gave me the opportunity to move from software development to NGS analysis. I greatly admire his discipline, integrity and passion for science.

I cannot say thank you enough to my mother and husband for being the biggest support for pursuing PhD and for their immense love and care. Gaurav for being the rock of emotional support, for all his patience and never letting me give up. Words will never be enough, Gaurav! Thanks to my dear father for all his love, support and for always saying that he lives for us.

I would also like to thank all members of the Spillane lab and Bioinformatics group for all their help and support. I would like to specially thank Peter McKeown and Sandesh Rao for always being there to discuss problems, helpful suggestions and checking my thesis. A special thanks to Paul Korir, Peter Keane and Martin Braud for helpful bioinformatics discussions and suggestions.

My sincere thanks to my parents-in-law for all their love and support. A special thanks to Sahil, Shweta and Isha for lots of talks and cheering me up. Thanks to my friends from Hyderabad, Preeti, Shelly and Rachit, who were never really away. My PhD would not be possible without your love, support and friendship. Thanks to Vineeta and Eva for their friendship, laughs and long chats. A big thanks to Mammor and Mags for making my stay wonderful in Ireland.
Summary

A range of different genetic processes can facilitate species-specific adaptations. These include orphan genes and genes subject to genomic imprinting. All organisms are known to contain unique genes that have no recognizable homolog in other lineages. These genes are known as orphan genes. Genomic imprinting refers to a phenomenon independent of Mendelian genetics where only one allele is actively expressed depending on its parent of origin. Both processes have been argued to play key roles in plant evolution but their molecular and evolutionary biology is poorly characterized, largely due to the fact that both orphan genes and imprinted genes have only been identified from a handful of model plant species.

In this thesis, I have identified orphan genes in the genomes of two crop species, *Camelina sativa* and *Sorghum bicolor*. *Camelina sativa* is an oilseed crop of the Brassicaceae family, whereas *Sorghum bicolor* is the fifth most important cereal crop belonging to the Poaceae family. Orphan genes identified in these two important crops are important sets of genes that may have played a role in acquiring adaptation to different environmental conditions and other unique characteristics to these species. In addition, a set of novel chimeric open reading frames (ORFs) is identified in *Cajanus cajan* which is another important food crop. The cytoplasmic male sterility (CMS) trait in plants is often associated with chimeric ORFs that are the products of mitochondrial genome rearrangements and can cause pollen abortion. The novel chimeric ORFs identified represent the most promising candidates for CMS-related mitochondrial rearrangements in *Cajanus cajan*. Therefore, my research on newly originated genes in agronomically important crops not only has meaningful evolutionary implications for the fundamental biology of species but can also assist plant breeding.

To further understand the evolution of plant imprinted genes and critically assess theories concerning their evolution and origins, I assessed the level of positive selection in orthologs of known *Arabidopsis thaliana* imprinted genes. Orthologs of imprinted genes were extracted from an orthology database developed in this Thesis from 34 sequenced Viridiplantae species. The results suggested a significant elevated
level of *Arabidopsis thaliana*-specific positive selection in *Arabidopsis thaliana* imprinted paternally expressed genes (iPEGs). Most of the positively selected sites identified in imprinted genes showed fixation in *Arabidopsis thaliana* population data. As positive selection may drive the fixation of beneficial traits within a population, imprinted genes that showed fixation of positively selected sites are the best candidates for further functional studies. Genomic imprinting in plants is identified to be mainly restricted to the seed endosperm. Genome-wide allele-specific expression analysis conducted in this thesis using a F1 triploid hybrid system demonstrates that altering the genome dosage can induce genomic imprinting in plant tissues (in embryos) where imprinting does not occur at the diploid level.
Chapter 1. Introduction
In this chapter, I provide general information about the evolution of new genes, duplication and selection. I also review the literature on orphan genes and imprinted genes.

1.1 Origination of new genes

Mutation and recombination are two major forces of evolution. Few of the first examples of evolution are emergence of RNA as first genetic molecules, conversion of RNA to DNA, emergence of first cells with DNA and linking together of genes into first chromosomes. Genes are gained and lost over the course of evolution. The number of genes can vary greatly between different species (Ranz and Parsch, 2012), ranging from 110 genes in Candidatus Tremblaya princeps (Lopez-Madrigal et al., 2011) to 89,418 genes in recent sequenced species Camelina sativa (Kagale et al., 2014). Such divergence can also be observed among closely related species. For example, Arabidopsis thaliana and Camelina sativa shared a common ancestor about 17 million years ago (mya) (Kagale et al., 2014). The total number of genes in Camelina sativa is approximately three times that of Arabidopsis thaliana. This variation in number of genes can be one of the explanations for the enormous diversity among different species. Therefore, the origin of new genes is considered as one of the important processes of evolutionary innovation, which contribute to species-specific specialism in different species (Arendsee et al., 2014; Neukermans et al., 2015).

The increasing availability of complete genome sequences from different species has made it possible to answer the important questions regarding the mechanism of origin and evolution of new genes. As the French biologist François Jacob famously wrote, “the probability that a functional protein would appear de novo by random association of amino acids is practically zero” (Jacob, 1977). Until the past few years it was a generally held view that new genes originate from old coding genes via duplication. Albeit it might seem improbable that new genes can arise de novo from non-coding DNA, a number of studies have demonstrated that each genome contains a significant proportion of genes that arise de novo in the genome (Knowles and McLysaght, 2009; Wu et al., 2011; Neme and Tautz, 2013; Wu and Zhang, 2013; Andersson et al., 2015).
1.1.1 Mechanisms of origination of new genes

DNA-based duplication is the most widely studied mechanism of evolution of new genes (described in detail in section 1.2). However, duplicate gene copies may also arise via retrotransposition. Retrotransposition is a RNA mediated duplication mechanism whereby mRNA of parental gene is reverse transcribed and integrate at a new locus that leads to creation of new genes (Brosius, 1991; Kaessmann et al., 2009). It is also possible that new genes can arise via fusion of existing genes to form chimeric genes (Long and Langley, 1993; Liu et al., 2009; Ranz and Parsch, 2012; Sanchez-Puerta et al., 2015). Interestingly, new protein-coding genes can also arise de novo from non-coding DNA (Donoghue et al., 2011; Ranz and Parsch, 2012; Reinhardt et al., 2013; Silveira et al., 2013; Arendsee et al., 2014; Zhao et al., 2014; Andersson et al., 2015) or within an existing gene by making use of an alternative reading frame (Donoghue et al., 2011; Sabath et al., 2012; Neme and Tautz, 2013). Figure 1.1 represents the major mechanisms that contributes to the origination of new genes (Ranz and Parsch, 2012). Although de novo generation of new genes was previously thought to be rare, a recent study in yeast proposed that de novo gene birth could be more prevalent than duplication (Carvunis et al., 2012). A schematic life cycle of genes is suggested by Neme and Tautz (2014), that includes a phase where new genes can emerge from non-genic sequences (Figure 1.2).

Other mechanisms that can contribute to gene birth are transposon exaptation, exon shuffling and horizontal gene transfer. Formation of new genes through transposon exaptation results from co-opting or exaptation of transposable elements into ORFs. A total of 7.8% A. thaliana genes were identified to contain a region with close similarity to a known transposable element (TE) sequence (Lockton and Gaut, 2009) which also supports the hypothesis that TEs can be an important factor in lineage divergence (Oliver et al., 2013). Exon shuffling is another important mechanism of evolution of new genes where new genes are created by recombining already-existing protein-coding domains. In Drosophila, domain rearrangements occur in 35.9% of gene families (Wu et al., 2012). Lastly, Horizontal Gene Transfer (HGT) is the exchange of genetic material across normal mating barriers between different species.
(Keeling and Palmer, 2008). HGT has played a major role in evolution of bacterial species (Ochman, 2001).

Figure 1.1: Mechanisms of origin of new genes. a). Duplication and divergence mechanism where GeneA duplicates and the newly duplicated copy undergo divergence and become GeneB, b). Retrotransposition where mRNA of a gene is reverse transcribed and get integrated into the genome forming GeneB, c). Gene fusion where two genes fuse together to form a chimeric gene, d). De novo origination of GeneC from non-coding region.
Figure 1.2: Schematic representation of life cycle of genes. Blue arrows represent transitions which lead, either partially or completely, to a new gene, and are therefore dubbed processes of gene birth. Red arrows represent the loss of features which result in the degradation of the genic potential of a sequence. Green arrows represent the processes which increase the gene repertoire from existing genes. Raw material for genes is stochastically generated as protogenes, entities that have gene-like properties, but may still lack a proper function. Once a protogene is able to perform a function that has an adaptive advantage, it will become fixed in a population. Gene loss through pseudogenization can lead to the death of a gene in a lineage, when the selective pressure upon it is released. This will also be the case for protogenes, which have not fully developed into genes (Source: Neme and Tautz, 2014).
1.2 Duplication and origin of new genes

Gene duplication provides fuel for two major engines of evolution: adaptation and speciation. In 1970 Susumu Ohno proposed that gene duplication events were a major force behind genome evolution, giving rise to novelty in the genome (Ohno, 1970). Sturtevant (1925) initially suggested the importance of gene duplication in generating new genes (Sturtevant, 1925). Further similar studies later supported this notion that duplicate genes are an important source for creating new genes (reviewed in (Nei, 2013)). It has also been suggested that in eukaryotic species, duplicate genes arise at a very high rate, on average 0.01 gene per million years (Lynch and Conery, 2000).

1.2.1 Generation of duplicate genes

There are several ways duplication can contribute to the evolution of a species:

- Duplication of its entire genome;
- Duplication of a single chromosome;
- Duplication of a single gene or group of genes.

1.2.1.1 Duplication of the entire genome

Duplication of a whole genome (also referred to as whole-genome duplications (WGDs)), gives birth to polyploid organisms and is considered as a major force in the evolution of angiosperms, among others. WGD is considered as the most effective mechanism for increasing the number of genes in the genome. However, a WGD does not necessarily double the number of genes as newly arisen genes are mostly silenced or rapidly eliminated from the genome (Adams and Wendel, 2005; Scannell et al., 2007). For instance, in the case of yeast, 92% of all genes could have been lost since a genome duplication event that occurred 100 mya (Seoighe and Wolfe, 1998).

WGD has been demonstrated to contribute to increased diversification, speciation, adaptive radiation, and genomic novelty (Levin, 1983; Soltis et al., 2004; Rieseberg and Willis, 2007; Soltis et al., 2009; Fawcett and Van de Peer, 2010; Bomblies and Madlung, 2014). Analysis in angiosperms supported the hypothesis that polyploidy leads to a dramatic increase in species richness (Soltis et al., 2009).
All the major clades of angiosperm lineages are reported to have undergone rounds of WGDs, subsequent and/or repeated, during their course of evolution (Figure 1.3) (Soltis et al., 2009; Jiao et al., 2011; Garsmeur et al., 2014; Jiao et al., 2014; Vanneste et al., 2014). For example, the model plant species *A. thaliana* (and probably all the core Brassicaceae genera) has experienced three ancient whole-genome duplication (WGD) events: alpha, beta and gamma (Bowers et al., 2003) (Figure 1.3).

**Figure 1.3:** Representation of the putative times of ancestral polyploidy events inferred for flowering plants. *Ovals indicate the genome duplications identified in sequenced genomes. The diamond refers to the triplication event probably shared by all core eudicots. Horizontal bars denote confidence regions for ancestral seed plant WGD and ancestral angiosperm WGD (Source: Jiao et al., 2011).*

The alpha duplication event pre-dated the diversification of core Brassicaceae from *Aethionema*, and occurred about 40 mya (Franzke et al., 2011). The beta event pre-dated the divergence of *Arabidopsis* from other core eudicots, but post-dated the divergence from the monocots about 70 mya. The gamma event was linked to the
diversification of eudicots and perhaps all angiosperms. Figure 1.4 represents the phylogenetic tree of sequenced plant and algae species and their major clades. The 34 sequenced species (that is, the number of species sequenced in 2011) represented in Figure 1.4 were used to develop an orthology database, which I described in Chapter 4.

Figure 1.4: Phylogeny of sequenced plant and green algal genomes. Major plant and green algal clades are labeled in red (figure modified from phytozome.org).
1.2.1.2 Duplication of a single chromosome

The condition when the number of chromosomes is not an exact multiple of the monoploid number of that species is called aneuploidy. In mammals, abnormal chromosome numbers are lethal and regularly lead to spontaneous abortions (miscarriages), or result in a genetic disease like Down syndrome. Plant genomes are considered to have a better system to tolerate aneuploidy than animals. However, even in plants aneuploid genomes are not entirely stable. As a general rule, aneuploidy has a much more detrimental effect on the phenotype than changes in dosage of the whole genome (Birchler, 2013). Considering the genomic instability and lethal effects, aneuploids can be excluded from the major factors that contributed to the evolution.

1.2.1.3 Duplication of single gene or group of genes

Gene families may evolve through segmental duplications (including local tandem duplications) or gene transposition. The increase of gene number by segmental duplications is generally small, but it may produce thousands of genes if it occurs repeatedly. One such example is olfactory receptor genes in vertebrates (Nei, 2013). Transposition of genes from one genomic location to another occurs via transposons and can provide hundreds of thousands of copies.

1.2.2 Evolutionary fate of duplicate genes

Similar to a point mutation, duplication can be fixed or lost in the population. The long-term evolutionary fate of the duplicated genes that become fixed is determined by the functions that newly duplicated copies acquire. Three alternative outcomes have been suggested in the evolution of duplicate genes: (i) one copy may become silenced (non-functionalization); (ii) one copy may acquire a novel function and be preserved by natural selection, while the other copy retains the original function (neo-functionalization); (iii) both copies may become partially compromised by mutation accumulation and their total capacity is reduced to the level of the single-copy ancestral gene (sub-functionalization) (Lynch and Conery, 2000).
A classic example of sub-functionalization is the MADS-box family, where two duplicated genes, ZMM2 and ZAG1, share the ancestral AGAMOUS gene function in maize (Mena et al., 1996). AGAMOUS gene in Arabidopsis is required for male and female reproductive organ development and floral determinacy. Duplicated genes ZAG1 and ZMM2 has overlapping but non identical activities in maize (Mena et al., 1996). OEP16, a channel protein, is an example of neo-functionalization (Drea et al., 2006), where both the copies (OEP16-L and OEP16-S) diverged in terms of expression patterns and coding sequence. He and Zhang (2005) suggested an interesting possibility of subneofunctionalization, where they discovered a combination of neo-functionalization and sub-functionalization. Using protein-protein interaction data from yeast, the analysis demonstrated rapid sub-functionalization, accompanied by prolonged and substantial neo-functionalization in a large proportion of duplicate genes (He and Zhang, 2005). The authors also argued that the pure sub-functionalization or neo-functionalization models are not adequate to explain the genomic patterns of protein interaction or gene expression for duplicate genes.

Different types of genes may vary in terms of their mutation robustness. Some genes require several substitutions to acquire a new function, while others require a very few. For example, genes of the terpene synthase family in Picea abies have undergone multiple rounds of neofunctionalization (Keeling and Palmer, 2008). In contrast, LEAFY transcription factors seems to lack functional plasticity: they are retained as a single copy in most angiosperms and after duplication there is a strong evidence of neutral mutation occurring (Baum et al., 2005).
1.3 Orthologs, paralogs and their functional conservation

Orthologs and paralogs, are both kinds of homologous gene, originated through speciation and duplication, respectively. Walter Fitch introduced the concept of orthology and paralogy more than four decades ago (Fitch, 1970), and gave the most widely accepted definition of orthologs. As per the original definition, orthologs (‘ortho’ means exact) are defined as genes from different species that derive from a single gene in the last common ancestor. Paralogs (‘para’ means parallel) genes are derived from a single gene that was duplicated within a genome (Fitch, 1970, 2000).

Although the original definitions of orthologs and paralogs are very simple to understand, but when multiple species and multiple genes are considered these definitions are complicated by combinations of gene losses, duplications and horizontal transfer events. Additionally, the most common features of eukaryotic genomes, such as complex domain architecture, alternative splicing and alternative transcription further complicated these concepts (Gabaldon and Koonin, 2013). This complexity has lead to the adoption of additional definitions that make it easier to refer to the type of evolutionary event when comparing multiple genes and species. Figure 1.5 explains the concepts of orthologs, co-orthologs, inparalogs and outparalogs using an arbitrary tree. Co-orthologs, which are also called ‘lineage-specific expansion of paralogs’ are paralogs produced by duplication of orthologs subsequent to a speciation event. Inparalogs are paralogs in a given lineage that have evolved through a duplication event which happened after the speciation event, whereas, outparalogs are paralogs in a given lineage which have evolved by gene duplication that happened before the speciation event (Sonnhammer and Koonin, 2002).

The sequence and functional conservation of orthologs and paralogs has recently been in debate (Dessimoz et al., 2012). According to the ‘orthology conjecture’ hypothesis (or standard model of phylogenomics), orthologs are more likely to be functionally similar than paralogs (reviewed by (Koonin, 2005)). Nehrt et al. (2011) rejected this hypothesis by demonstrating that paralogs from the same species are often a much better predictor of functional divergence than are orthologs or paralogs from different species (Nehrt et al., 2011; Altenhoff et al., 2012).


**Figure 1.5**: Example tree to explain the concept of inparalogs, co-orthologs, outparalogs. Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human and worm lineages. In this case, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence ‘inparalogs’ to each other when comparing human to worm. By contrast, the genes HB and HA* are ‘outparalogs’ when comparing human with worm. However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication (Source: Sonnhammer and Koonin, 2002).
Further studies reported that the Gene Ontology (GO) annotations biases had impacted Nehrt et al. (2011) results (Altenhoff et al., 2012; Thomas et al., 2012). On removing these biases, it was found that orthologs actually have more similar functional annotations than paralogs, although with small but significant support for the ‘ortholog conjecture’ model (Altenhoff et al., 2012). Using RNA-seq data, different studies supported ‘ortholog conjecture’ by showing that orthologs display conserved expression patterns compared to paralogs (Huerta-Cepas et al., 2011; Chen and Zhang, 2012). Regarding this argument, experts in the latest ‘quest for orthologs’ consortium meeting so far concluded that orthology concept does not imply identical function (Sonnhammer et al., 2014). Therefore, it would be beneficial to estimate the functional conservation between two genes given their relationship rather than looking for functional conservation in orthologs. Further work on the conclusion on the concept of functional conservation of orthologs is in progress (Sonnhammer et al., 2014).

Identification of orthologs is generally the first step in any molecular evolution study. Therefore, accurate determination of orthologs is crucial for evolutionary and functional annotation studies. A plethora of standalone tools, databases and web-services has been developed for identification of orthologs (Overbeek et al., 1999; Remm et al., 2001; Li et al., 2003; Alexeyenko et al., 2006; Huerta-Cepas et al., 2008; Datta et al., 2009; Egan et al., 2009; Vilella et al., 2009; Lechner et al., 2011; Afrasiabi et al., 2013; Powell et al., 2014; Altenhoff et al., 2015; Kriventseva et al., 2015) and scientific meetings are devoted for the research on orthologs (Gabaldon et al., 2009). These tools are based on different algorithms and can be majorly divided in two categories: a) phylogeny-based tools and b) BLAST-based tools. Phylogenetic based methods are considered to be computationally more expensive than the BLAST-based methods, and often contains errors, especially at large evolutionary distances (Kristensen et al., 2011). There is a continuous debate on the best method/tool for prediction of orthologs and several benchmarking studies have assessed the performance of these methods (Chen et al., 2007; Datta et al., 2009; Salichos and Rokas, 2011). Figure 1.6 shows one such comparison of different tools as an example (Datta et al., 2009).
Figure 1.6: Results of different orthology prediction methods assessed using a benchmark dataset of 100 genes from the TreeFam-A resource. Methods evaluated include OrthoMCL-DB, InParanoid, SCI-PHY and PHOG variants. Recall measures the fraction of TreeFam-A orthologs detected by a method. Precision measures the fraction of method’s predicted orthologs that are included in TreeFam-A (Source: Datta et al., 2009).

Due to the lack of an orthology database for all sequenced plant species, I have built a plant orthology database to identify orthologs groups in 34 sequenced Viridiplantae species. Chapter 4 describes in detail the identification of orthologs groups in the protein-coding sequences of sequenced Viridiplantae species using the orthology prediction program OrthoMCL (Li et al., 2003).
1.4 Orphan genes and *De novo* genes

The term ‘orphan’ was introduced by Bernard Dujon in 1996 to reflect the discovery of a large number of genes (about 26%) in the yeast genome that showed no homologs in any other species (Dujon, 1996). Orphan genes can be defined as genes with coding sequences unique to the species. Fisher and Eisenberg (1999) identified the orphan genes in all sequenced microbial species and confirmed that the presence of orphan genes in each species is not an artifact of annotation or sequencing. They also introduced the term ORFans (orphan ORFs) (Fischer and Eisenberg, 1999). Orphan genes can be defined within a specific taxonomic group or lineage and can also be called taxonomically restricted genes (TRGs) (or lineage-specific genes) (Wilson et al., 2005). Genome sequencing initiatives have revealed that orphan genes exist in all biological lineages of life. The number of orphan genes in various species ranges wildly from <1–71% with 5–15% being fairly typical (Arendsee et al., 2014).

Orphan genes can constitute a part of the dispensable genome in the ‘pan genome concept’ (Figure 1.7). Part of the genome that is present in all the individuals of a species is termed the ‘core genome’ and part of the genome that is shared by a subset of individuals is termed the ‘dispensable genome’. The sum of the core genome and dispensable genome is the total genomic content for that species, i.e. the pan genome. The pan genome concept was initially developed for bacterial species due to the relative simplicity of their genomes compared to eukaryote genomes (Tettelin et al., 2005) and was later extended to plants (Morgante et al., 2007; Hirsch et al., 2014).

Different studies conducted on orphan genes identified general trends in the sequence features of orphan genes across different species. These general genomic characteristics include: short length, fewer introns, atypical GC content, and increased evolutionary rates (Domazet-Loso and Tautz, 2003; Cai et al., 2006; Wilson et al., 2007; Donoghue et al., 2011; Palmieri et al., 2014). Enrichment for cysteine-rich peptides was also identified to be a characteristic of orphan genes (Donoghue et al., 2011; Arendsee et al., 2014).
We now have a larger number of fully sequenced genomes from various taxa, which in some cases has made it possible to accurately determine the origin and age of orphan genes within an appropriate phylogenetic framework. For example, founder genes (the orphan genes that gave rise to the new genes in descendant lineages) can be allocated to certain time points in evolution. Domazet-Loso et al. (2007) proposed a general ‘Phylostratigraphy’ approach that can be used to systematically identify all orphan genes within the evolutionary lineages that have led to a particular extant genome.

Orphan genes can arise in a lineage through a range of molecular evolutionary processes. There are two main models to investigate how founder genes emerge in a genome: duplication-divergence and de novo emergence of new genes. The duplication-divergence model proposes that a new gene would initially be created through a gene duplication or transposition event (Figure 1.1) and would then go through a phase of fast adaptive evolution, during which time it would lose all similarity to its parental gene (Domazet-Loso and Tautz, 2003). These duplication events can occur through recombination-mediated events or virus/transposon-mediated processes. Duplication and divergence has long been thought as the primary mechanism for the evolution of new genes (Tautz and Domazet-Loso, 2011). Duplication of non-orphans genes was found to be responsible for the evolutionary
origins of almost one quarter of all the putative Brassicaceae-specific orphans in A. thaliana genome (Donoghue et al., 2011). Donoghue et al. (2011) also concluded that the generation of an orphan gene via rapid evolution is more likely when the parental duplicate gene is already experiencing elevated rates of evolution.

A subset of orphan genes can directly originate as de novo genes from non-coding sequences (Donoghue et al., 2011; Tautz and Domazet-Loso, 2011; Carvunis et al., 2012; Xie et al., 2012; Abrusan, 2013; Silveira et al., 2013; Zhao et al., 2014). Recent studies have provided support for the functional relevance of de novo gene birth processes to the generation of evolutionary novelty (Carvunis et al., 2012; Abrusan, 2013; Neme and Tautz, 2013; Arendsee et al., 2014; Neme and Tautz, 2014). In addition to originating directly from non-coding DNA, de novo gene emergence can occur via overprinting. Overprinting describes a mechanism when a novel open reading frame originates through point mutations inside an existing gene (Neme and Tautz, 2013), which leads to the expression of a novel protein in an alternative reading frame of an existing gene. These genes are important to study as they retain the parental gene, which excludes the possibility of misidentifying a highly diverged gene, or a gene transferred from other organisms (e.g. by horizontal gene transfer) as a newly evolved gene. Wadhawan et al. (2008) identified an interesting case of overprinting (Gnas and Gnal complex) where the overlap between coding regions not only results in the production of two unrelated proteins but also an elaborate pattern of imprinting, alternative splicing, and antisense transcription between the two loci (Wadhawan et al., 2008). Neme and Tautz (2013) used a phylostratigraphy approach and identified cases of overprinted genes among annotated genes in mouse genome, where the alternative reading frame maps to a different phylostratum (age group) than the original reading frame. This study has suggested that existing genes might readily become templates for de novo evolution of new functions within them (Neme and Tautz, 2013). However there is no such comprehensive study in plants although we have a number of sequenced plant species available in public domain (http://www.phytozome.net/). Considering the importance of these genes, I have identified overprinted genes in A. thaliana (Chapter 2).

When the emergence of an orphan gene in a lineage is followed by the acquisition of a biological function, such orphan genes have the potential to provide lineage-specific
adaptations (Domazet-Loso and Tautz, 2003; Guo et al., 2007; Donoghue et al., 2011; Arendsee et al., 2014; Palmieri et al., 2014; Neukermans et al., 2015). For example, in *Drosophila* some *de novo* evolved genes have been demonstrated to be essential (Chen et al., 2010; Reinhardt et al., 2013). The *QQS* gene of *A. thaliana* is the first orphan gene with a biochemically-characterized function (reviewed in (Arendsee et al., 2014)). This gene acts as a regulator of carbon and nitrogen allocation, affecting carbon and nitrogen partitioning to starch, lipid and protein in leaves and seeds.

Determining the functionality of orphan genes is challenging, as the usual *in silico* homology-based prediction methods are not possible (or are very limited). Another *in silico* approach is to identify the expression patterns for these genes. As orphan genes are considered to potentially provide lineage-specific adaptations, it would be expected that their function and expression could be related to specific developmental stages or environmental conditions. Donoghue et al. (2011) demonstrated that *A. thaliana* orphan genes are enriched for genes responsive to a range of abiotic stresses. Similarly, orphan genes were also identified to be abiotic stress responsive in yeast (Carvunis et al., 2012), rice (Guo et al., 2007) and water flea (Colbourne et al., 2011). In addition to stress, roles in reproduction have also been demonstrated for orphan genes. In *Drosophila* and mouse, orphan genes are highly overrepresented in testes (Neme and Tautz, 2013; Arendsee et al., 2014) while *QQS* is an example of a plant orphan gene that is highly expressed in pollen (Wang et al., 2008; Li et al., 2009). These studies highlight the functional significance of orphan genes and support the notion that orphan genes play a significant role in providing lineage-specific adaptation.
1.5 Fate of newly arisen mutations

The evolutionary fate of newly evolved mutations is determined by random genetic drift, positive selection or purifying selection. Comparing the relative fixed rates of non-synonymous (amino acid altering mutations) and synonymous (silent mutations) changes provides an important strategy to study the mechanisms of selection sequence evolution. Synonymous mutation does not change the amino acid encoded whereas nonsynonymous mutations lead to a change in the amino acid that can alter the protein folding, or interfere with binding to other molecules, subsequently alter or destroy protein function. Therefore the substitution rate of the nonsynonymous mutations ($dN$) is function of selection pressure on the protein product of a gene while substitution rate of the synonymous mutations is considered to be neutral. The ratio of nonsynonymous substitutions per nonsynonymous site ($dN$) to synonymous substitutions per synonymous site ($dS$) is known as omega ($\omega$), and is an indication of selection pressure. If an amino acid change is selectively neutral, it will be fixed at the same rate as a synonymous mutation i.e. $dN=dS$ and $\omega = 1$. Therefore, $\omega > 1$ signifies positive selective pressure, $\omega = 1$ signifies neutral evolution, while $\omega < 1$ indicates purifying selective pressure. An $\omega > 1$ means that nonsynonymous mutations offer fitness advantages to the individual and have higher fixation probabilities than synonymous mutations (Yang and Bielawski, 2000). A few recent studies have experimentally demonstrated the effect of positive selection on protein function (Tennessen, 2008; Moury and Simon, 2011; Loughran et al., 2012). Loughran et al. (2012) used site-directed mutagenesis to replace positively selected residue with the ancestral amino acid in the human myeloperoxidase enzyme and demonstrated a link between positive selection and a functional shift to a novel peroxidase activity.

Although high levels of positive selection are evident in various species, it is a general view that most protein coding genes have evolved under purifying selection (Li, 1997; Hughes, 1999). A high proportion of amino acids, in most protein coding genes, are largely invariable due to strong functional constraints. Gossmann et al. (2010) found a little evidence of positive selection in plants when conducting a genome wide analysis across different plant species (Gossmann et al., 2010), consistent with previous
studies done in plants (Bustamante et al., 2002; Barrier et al., 2003; Schmid et al., 2005).

1.5.1 Methods for identifying selection pressures acting on protein-coding sequences

Until 2000, a number of methods were suggested to calculate \( dN \) and \( dS \) between two protein coding sequences (Miyata and Yasunaga, 1980; Li et al., 1985; Nei and Gojobori, 1986; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995; Yang and Bielawski, 2000). These models were described as approximate methods, as all made simple assumptions about the nucleotide substitution process (reviewed in (Yang and Bielawski, 2000)). These methods used the following steps for the calculations: counting synonymous (S) and nonsynonymous (N) sites in the two sequences, counting synonymous and nonsynonymous differences between the two sequences, and correcting for multiple substitutions at the same site. Here S and N are sequence length multiplied by the proportions of synonymous and nonsynonymous changes. Yang and Nielsen (2000) improved the approximate methods by taking into account two major features of nucleotide sequence evolution: transition/transversion rates and codon-usage biases. Furthermore, for approximate methods, a normal approximation is applied to \( dN-dS \). Here, statistical tests can be used to test whether \( dN \) is significantly greater than \( dS \).

Another popular approach for revealing trends in synonymous and nonsynonymous variation is the sliding windows approach (Endo et al., 1996; Hurst and Pal, 2001; Fares et al., 2002). Schmid and Yang (2008) demonstrated the artifactual effect of sliding-window analysis through a re-analysis of the breast-cancer gene \textit{BRCA1} (Schmid and Yang, 2008). Another class of selection-measuring methods is the maximum likelihood (ML) method, which is based on codon substitution models. For ML, a likelihood ratio test (LRT) can be used. The null model assumes \( \omega \) fixed at 1, whereas the alternative model estimates \( \omega \) as a free parameter. The likelihood difference between the two models is compared with \( x^2 \) distribution with one degree of freedom to test whether \( \omega \) is different from 1. These methods are considered to be more robust than other methods of detecting positive selection like sliding windows approaches (Schmid and Yang, 2008). The following section describes the maximum likelihood methods in detail.
1.5.1.1 Maximum likelihood method for measuring selection pressure

Markov process models of codon substitution were developed to account for different \( \omega \) ratios among branches in the tree (Yang, 1998). The lineage of interest can be assumed to have different \( \omega \) ratio from the background \( \omega \) ratio for all other lineages in the phylogenetic tree. This two-ratio model can be compared with a model that assumes same \( \omega \) for all the lineages in the phylogeny (one-ratio model) to examine whether the lineage of interest has a different \( \omega \) ratio than other lineages. Figure 1.8 explains a model of variable \( \omega \) ratios on the branches of an example tree of four sequences. Nielsen and Yang (1998) implemented simple codon-substitution models that allowed inferring heterogeneous \( \omega \) ratios among sites. The neutral model assumes two classes of sites: conserved sites (\( \omega = 0 \)) and neutral sites (\( \omega = 1 \)). The selection model adds a third class of sites where the \( \omega \) ratios are estimated from the data. However, this model showed several limitations and was too simple to capture the substitution process of various proteins.

![Diagram of a phylogenetic tree with variable \( \omega \) ratios](image)

**Figure 1.8**: Explains a model of variable \( \omega \) ratios among branches of a tree of four sequences. The \( \omega \) ratio for the branch leading to species X1 (\( \omega_1 \)) is different from the ratio for all the other branches (\( \omega_0 \)).
The null model, M0, calculates the value of $\omega$ over the entire alignment. The M0 model assumes that all sites and all lineages are evolving at the same rate. Model M3 is an extension of M0 model and allows the value of $\omega$ to vary freely. The M3 model has two classes: m3 (k=2) discrete, which allows two variable site classes, and m3 (k=3) which allows three variable site classes. M1a is a neutral model that allows two site classes for proportion of sites where $\omega_0=0$ and $\omega_1=1$. M2a is a selection model that allows three site classes where $\omega_0=0$ and $\omega_1=1$ and $\omega_2$ is estimated. M1a and M2a are slight modifications of models M1 and M2 in Nielsen and Yang (1998), which had $\omega_0=0$ fixed and were found to be unrealistic for most data sets. The other test compares the null model M7, which assumes a $\beta$ distribution for $\omega$ and the alternative model M8 ($\beta$ and $\omega$), which adds an extra site class of positive selection. M8 is a parameter rich model, which contains 10 different $\omega$ site classes and an additional parameter for $\omega$ that is free to vary between 0 and >1. M8a is the null hypothesis of M8 where the 11th category is neutral, i.e. $\omega=1$.

Yang and Nelson (2002) further extended the codon substitution models to identify selection pressure acting on different sites along specified lineages. The development of these models were based on the idea that if positive selection acts on few amino acids at few time points then both the lineage-specific and site-specific models might lack the power to detect it. The new models allowed the $\omega$ ratio to vary both among sites and among lineages and are termed as ‘branch-site models’ (Yang and Nielsen, 2002). Branches that are tested for positive selection are termed the “foreground” branches and all others are considered “background” branches. Two versions of this model were implemented by the authors, termed model A and model B. Model A assumes that two site classes are the same in both foreground and background lineages ($\omega_0=0$ and $\omega_1=1$) and $\omega_2$ is calculated. In model B, both $\omega_0$ and $\omega_1$ are estimated from the data as free parameters. Simulation studies showed that these models were sensitive to model assumptions, however, model A performed marginally better than the model B (Zhang, 2004). Zhang et al. (2005) later introduced modifications to the branch-site models, where two additional site classes allowing for positive selection were added to model A. Model A null is the null hypothesis for this model that allows for sites evolving under purifying selection, or neutrally evolving in the background lineages. These improved models outperformed the previous branch-site models developed by Yang and Nielsen (2002). Table 1.1.
shows the comparison of codeml models for the identification of positive selection, degrees of freedom (df), the difference between the models and the critical $\chi^2$ values (modified from Morgan et al. 2010).

Table 1.1: Calculation for Likelihood Ratio test (LRT).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>df</th>
<th>$\Delta l$</th>
<th>Critical $\chi^2$ values</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0 vs M3k2</td>
<td>2</td>
<td>X2</td>
<td>$\geq 5.99$</td>
</tr>
<tr>
<td>M3k2 vs M3k3</td>
<td>-</td>
<td>X1</td>
<td>$\geq 1.00$</td>
</tr>
<tr>
<td>M1a vs M2a</td>
<td>2</td>
<td>X2</td>
<td>$\geq 5.99$</td>
</tr>
<tr>
<td>M7 vs M8</td>
<td>2</td>
<td>X2</td>
<td>$\geq 5.99$</td>
</tr>
<tr>
<td>M8 vs M8a</td>
<td>1</td>
<td>X2</td>
<td>$\geq 2.71$ (@5%)</td>
</tr>
<tr>
<td>M1a vs Model A</td>
<td>2</td>
<td>X2</td>
<td>$\geq 5.99$</td>
</tr>
<tr>
<td>Model A vs Model A null</td>
<td>1</td>
<td>X2</td>
<td>$\geq 3.84$ (@5%)</td>
</tr>
</tbody>
</table>

After ML estimates of the model parameters are obtained, it is possible to use an empirical Bayes approach to infer the most likely site classes for each site. When the $\omega$ ratio for some sites is greater than 1, the Empirical Bayes (EB) approach can be used to calculate the posterior probabilities that each site belongs to the site class of positive selection on the foreground lineages. Yang and Neilsen (2002) used the Naive empirical Bayes (NEB) approach, which uses maximum likelihood estimates (MLEs) of parameters but does not take any sampling errors into account. Yang et al. (2005) proposed a Bayes empirical Bayes (BEB) procedure, which is a more robust approach than NEB. The BEB approach reduces the rate of false positives when analyzing small datasets and retains the power of NEB when analyzing large datasets (Yang et al., 2005).

In Chapter 5, I have used codeml models to identify the signatures of positive selection on A. thaliana imprinted genes. Population genetics and phylogenetics methods for detecting selection are complementary tools for evolutionary analysis. Population genetics tools are suitable for detecting recent selection, while phylogenetic tools suit the analysis of data from longer evolutionary periods. I have
taken advantage of the availability of both plant interspecific data (34 Viridiplantae species from the Phytozome database (Goodstein et al., 2012)) and the sequencing of 80 accessions of *A. thaliana* (Wang et al., 2012). *A. thaliana* population data was used to identify whether the positively selected sites identified in interspecies data are fixed among different populations of the *A. thaliana* species.
1.6 Dosage Effects

The concept of gene dosage balance arose with the work by Blakeslee et al. (1920) in the flowering plant species *Datura stramonium* and Bridges (1925) in *Drosophila melanogaster* (Blakeslee et al., 1920; Bridges, 1925). Addition of a single chromosome was found to be detrimental in both organisms, whereas polyploids were found to be viable (Birchler and Veitia, 2012). Further studies showed that the addition of a chromosome to a diploid has less severe effects than addition of a chromosome to a haploid. The relative dosage of chromosomal segments was also identified to be critical for normal development and phenotype (reviewed in (Birchler et al., 2001; Birchler and Veitia, 2012)). In both plants and animals, an increase or decrease in chromosomal dosage have a significant impact on the phenotype (reviewed in (Birchler et al., 2001)). Although in plants aneuploid genomes are not entirely stable, but in general, plants tolerate extra chromosomes better than animals (Matzke et al., 2003).

Changing the dosage can increase (direct trans effect) or decrease (via inverse trans effects) the gene expression of a target gene as the dosage of the effective chromosomal segment increases (Figure 1.9). An inverse effect is the more common type in both maize and *Drosophila* (Birchler et al., 2001). Surprisingly, in the maize *Adh* (*Alcohol dehydrogenase-I*) gene, Birchler (1979) showed that the total amount of ADH present in a one to four dosage series was nearly equivalent to the diploid level. In other words, no dosage effect was identified for *Adh* gene. This ‘no dosage effect’ phenomenon was referred to as dosage compensation (reviewed in (Birchler and Veitia, 2007)).

In flowering plants, genomic imprinting is proposed to be a mechanism that can function in balancing the gene dosage of the triploid endosperm tissue (Birchler et al., 2001; Birchler et al., 2005; Tiwari et al., 2010; Zhang et al., 2011), in which imprinting is mainly observed. In other words, some genes are imprinted because of their sensitivity to gene dosage for proper endosperm function. However, gene dosage is only one of many hypotheses that have been suggested as possible explanations for the evolution of imprinting. In the following section, I discuss genomic imprinting and the evolutionary hypotheses, which attempt to explain it.
Figure 1.9: Representation of the types of effects observed for gene expression in aneuploids. The x-axis represents the chromosomal dosage. The y-axis represents the percentage of expression in the aneuploid compared with the diploid. (A) A gene dosage effect occurs when a gene produces a proportional amount of product to its copy number. (B) Direct transacting effects when a gene is modulated in expression in direct correlation with a different chromosomal dosage. (C) Inverse dosage effect occurs when the expression of a target gene is inversely correlated with the dosage of another chromosomal region. (D) Dosage compensation occurs when the expression of a structural gene does not change with its dosage (Source: Birchler and Veitia 2012).
1.6.1 Genomic imprinting

The laws of inheritance were defined by Mendel more than a century ago. A number of genetic phenomena have been observed in both plants and animals that do not follow the classical Mendelian inheritance, also termed as non-Mendelian inheritance (McKeown and Spillane, 2014). Genomic imprinting is one of the examples of non-Mendelian genetics, together with extranuclear inheritance, gene conversion, infectious heredity, mosaicism and trinucleotide repeat disorder. Diploid organisms receive two sets of genes from their parents: one from the father and one from the mother. They behave equally in the offspring, i.e., their transcription is not expected to be affected by whether they are inherited from male or female parent. This is the standard picture of Mendelian genetics that holds for most of the genes. However, for some genes only one of the two alleles is actively expressed. This phenomenon is called genomic imprinting and has been observed in mammals, plants and insects (McKeown et al., 2013). Imprinting is an epigenetic effect as there is no change in the DNA sequence of the allele, but there is a change in the allele’s expression state depending on whether it has been transmitted via a female or male gamete. Therefore, genomic imprinting represents one form of epigenetic inheritance where heritable changes in genome function can occur without a change in DNA sequence (Bird, 2007).

1.6.1.1 Imprinted genes in plants

In plants, genomic imprinting has mainly been observed in the endosperm and it is associated to seed development. Endosperm is an ephemeral triploid tissue derived after fertilization of the diploid central cell with a haploid sperm cell. Imprinting was first described at the maize r1 locus, which shows parent-specific differences in expression during the development of the endosperm of the kernel (Kermicle, 1970). The first imprinted gene identified at the molecular level in plants was MEDEA in A. thaliana, which is an imprinted Maternally Expressed Gene (iMEG) essential for seed development (Kinoshita et al., 1999). In the ten years after the discovery of this first imprinted gene MEDEA a number of others were discovered in A. thaliana– namely, PHERES1 (PHE1) (Kohler et al., 2005), FLOWERING WAGENINGEN, FERTILIZATION INDEPENDENT SEED 2 (FIS2) (Jullien et al., 2006),
MATERNALLY EXPRESSED PAB C-TERMINAL (MPC) (Tiwari et al., 2008), ARABIDOPSIS FORMIN HOMOLOGUE 5 (AtFH5) (Fitz Gerald et al., 2009), AtMYB3R2, HOMEODOMAIN GLABROUS 3 (HDG3), HDG8, HDG9, and At5G62110, which encodes an unknown protein (Gehring et al., 2009; Kohler et al., 2012). Recent advances in sequencing technologies have made it possible to identify multiple imprinted genes including many nuclear proteins, such as transcription factors and chromatin related proteins in different plant species. A number of high-throughput sequencing screens and other genome-wide screens (Gehring et al., 2011; Hsieh et al., 2011; Luo et al.; McKeown et al., 2011; Wolff et al., 2011b; Zhang et al., 2011; Waters et al., 2013) have dramatically increased the number of candidate imprinted genes in A. thaliana, rice and maize. Surprisingly, a limited overlap is observed between A. thaliana imprinted genes identified by different high-throughput screens (Figure 1.10). One possible explanation for this lack of overlap could be the existence of high number of accession-specific imprinted genes (McKeown et al., 2013), as different high-throughput screen used the crosses between different accessions (Col-0 x Bur-0 (Wolff et al., 2011b) and Col-0 x Ler-0 (Gehring et al., 2011; Hsieh et al., 2011).

Apart from genome wide screens, a number of studies conducted for prediction of novel imprinted genes in mammals have demonstrated that machine learning strategies are also sophisticated method for identification of novel imprinted genes (Luedi et al., 2005; Luedi et al., 2007; Daura-Oller et al., 2009; Brideau et al., 2010). In 2005, Luedi et al. trained a statistical model based on DNA sequence characteristics that identified potentially imprinted in mouse genes and predicted the parental allele from which they are expressed. Following this, Luedi et al. (2007) developed a new algorithm for predicting the genome-wide imprint status of human genes directly from sequence features in the human genome. A new algorithm was developed for the human imprinting genes because in case of imprinting mapping predictions from one species to another by homology would result in significant error (Monk et al., 2006). For example, some genes are imprinted in both mouse and human, others including Igf2r, Ascl2, Tspan32, Cd81, Tssc4, Nap1l4, Gatm, Dcn, and Impact are imprinted in mouse but not human. A more robust approach for identification of imprinted genes was reported in 2010 where not only sequence features, but also epigenomic features were also taken into consideration for
prediction of novel imprinted genes (Brideau et al., 2010). This study has identified a total of 10 novel imprinted genes out of total 65 candidate genes predicted by the computational methodology. Though computational prediction of imprinted genes cannot substitute for experimental validation, but it can certainly provide a valuable set for identification of imprinted genes.

Figure 1.10: Imprinted genes in *A. thaliana* identified by different high-throughput screens showed limited overlap (Source: McKeown et al. 2013).

1.6.1.2 Evolution of genomic imprinting

A number of evolutionary hypotheses have been proposed to explain the evolution of genomic imprinting (Hurst and McVean, 1998), including dosage compensation (Iwasa, 1998), meiotic recombination (Pardo-Manuel de Villena et al., 2000), prevention of parthenogenesis (Varmuza and Mann, 1994), intralocus sexual conflict (Bonduriansky and Chenoweth, 2009) and the kinship hypothesis (Moore and Haig, 1991; Wilkins and Haig, 2003). Although the kinship hypothesis (which is commonly
called the parental conflict theory, PCT) has tended to be supported by studies in animals (Hurst and McVean, 1998; Wilkins and Haig, 2003), theories of imprinting remain under constant debate (Haig, 2013). Kinship hypotheses for imprinting evolution propose that imprinting has evolved because of an evolutionary conflict in individuals between maternally and paternally derived alleles. Selection due to the antagonistic co-evolution will derive mono-allelic expression of paternally derived alleles that increase maternal resource allocation to the offspring (growth enhancers), while growth inhibitors are predicted to be subject to mono-allelic expression from the maternally derived allele (Moore and Haig, 1991). However, the kinship theory has received support from the study of interploidy crosses in plants, and there is also some data to suggest that the endosperm is particularly sensitive to changes in the parental chromosome dosage (Jullien and Berger, 2010). Specifically, an increased dosage of paternal chromosomes promotes endosperm development while an increased dosage of maternal chromosomes repress endosperm development (Kohler and Weinhofer-Molisch, 2010).

If the parental conflict involves antagonistic co-evolution between imprinting regulators of offspring growth than the protein products of imprinted genes could be subject to rapid evolution via positive Darwinian selection (Hurst and McVean, 1998). The imprinted plant gene MEDEA, which originated during a whole-genome duplication, has undergone positive Darwinian selection after duplication which supports the kinship theory (Spillane et al., 2007). In contrast to the kinship theory mammalian imprinted genes provided a little evidence of the protein-coding regions of imprinted genes evolving under positive selection, though it remains possible that the regulatory regions of imprinted loci could be fast evolving in a manner that affects the timing, location or level of gene expression (O'Connell et al., 2010).
1.7 Aim of Thesis

An adaptation is a mutation or a genetic change that helps an organism to survive in its environment. Over the course of evolution, a number of genetic processes are evolved that can provide species-specific adaptations. Orphan genes and genes subject to genomic imprinting are two such processes. The aim of this thesis is to identify orphan genes in different plant species and assess theories concerning the evolution and origins of imprinted genes.

The objective of Chapter 2 is to identify orphan genes in two crop species, *Camelina sativa* and *Sorghum bicolor*. There is growing evidence that orphan genes play an important role in the environmental adaptation of a species. *Camelina sativa* is an oil seed crop that has the potential for biofuel applications due to its exceptionally high content of omega fatty acid. On the other hand, *Sorghum bicolor* is an important food source in semi-arid/arid regions and is one of the most drought resistance crops. Identification of orphan genes in such species provides us candidate genes that may have played a role in acquiring adaptation and other unique characteristics to both the species. Chimeric ORFs generated as a result of genome rearrangements are subset of orphan genes. The objective of Chapter 3 is to identify such chimeric ORFs in *Cajanus cajan* that are originated as a result of mitochondrial genome rearrangements. These chimeric ORFs can be considered as most promising candidates for cytoplasmic male sterility (CMS) in pigeon pea.

The objective of Chapter 5 and 6 is to understand the evolution of plant imprinted genes and assess the theories related to their evolution and origins. A number of imprinted genes are identified in *Arabidopsis thaliana* by various high-throughput screens and other studies. In Chapter 5, I assess the level of positive selection in orthologs of known *Arabidopsis thaliana* imprinted genes. The orthologs of imprinted genes are extracted from the plant orthology database developed in Chapter 4. The impact of positive selection can also be assessed at the population genomic level as variation subject to positive selection is expected to go to fixation throughout populations.
Genomic imprinting is mainly restricted to the seed endosperm in plants. However, the phenomenon of imprinting has been demonstrated to be disrupted by interploidy crosses that has led to an alternative hypothesis that the triploid genome in the endosperm might be required for imprinting in plants. Therefore in Chapter 6, a genome-wide allele-specific expression analysis is conducted to demonstrate that altering the genome dosage can induce genomic imprinted in plant tissues where imprinting is not observed at the diploid level.
Chapter 2. Orphan genes and *de novo* originated overprinted genes identified in plant species

Orphan genes identified in *Camelina sativa* are published as part of the *Camelina sativa* genome sequencing paper:

2.1 Introduction

Evolution of orphan genes is one of the genetic processes that have been argued to provide species-specific adaptations. Orphan genes are protein-coding genes that have no recognizable homolog in other species or within existing protein/nucleotide sequence databases. Defining a gene as an orphan gene involves the identification of all its potential homologous relationships. Homology between distantly related sequences could be difficult to identify due to mutations accumulated in low sequence similarity regions that can mask the homology signal. Therefore, compared to global search algorithms (e.g. ClustalW (Thompson et al., 1994) and Needleman-Wunsch (Needleman and Wunsch, 1970)), local search algorithms (e.g. BLAST (Altschul et al., 1990) and Smith-Waterman (Anderson et al., 1981)) are considered better for such comparisons as these avoid the low similarity regions altogether. In addition to local search methods, more sophisticated position-specific methods such as PSI-BLAST (Altschul et al., 1997) can be used to search for distant homology between sequences. These methods can recognize the distant similarities that are recognizable only through three-dimensional structures (Brenner et al., 1998).

Expectation score is used as a measure of significance of alignment. The expectation value (e-value) describes the number of hits one can expect to occur in a database search by chance. The lower the expectation value, the lower the probability that the alignment is merely a result of random chance. The choice of e-value cut-off can affect the significance of reported alignments that were used to define orphan genes. Considering the sensitivity of BLAST to e-value cutoff, Donoghue et al. (2011) showed the effect of different e-value cutoffs on the number of orphan genes identified in A. thaliana. An e-value threshold of 1e-2 was considered as the optimal threshold for identification of orphan genes in A. thaliana (Donoghue et al., 2011). The choice of e-value threshold of 1e-2 is also reflected in other studies for identification of orphan genes (Siew and Fischer, 2003a, b; Siew et al., 2004; Wilson et al., 2005). I have used the pipeline defined by Donoghue et al. (2011) for identification of orphan genes in Camelina sativa and Sorghum bicolor as part of two collaborative projects. Databases searched for defining the orphan genes include the fully sequenced plant species database Phytozome (Goodstein et al., 2012) and the NCBI sequence databases. PSI-BLAST (Altschul et al., 1997) and InterProScan
(Jones et al., 2014) were used to identify distant similarity and similarity with known domains respectively. Donoghue et al. (2011) demonstrated InterProScan to be less sensitive to e-value changes.

The orphan gene set identified in both the species further raised question regarding their evolutionary origins. Where do orphan genes come from, i.e. what evolutionary mechanisms are involved for origination of orphan genes? In A. thaliana, duplication of non-orphan genes was identified to be the mechanism of origination of a quarter of orphan genes while almost half of the orphan genes originated de novo (Donoghue et al., 2011). Interestingly, A. thaliana orphan genes also showed enrichment for overlapping genes. These results are consistent with the hypothesis that orphan genes can directly originate as de novo genes (Donoghue et al., 2011). De novo gene birth is poorly understood compared to duplication because it is considered improbable that new genes can arise de novo and was expected to produce polypeptides without any significant biological functions (Jacob, 1977; Siepel, 2009). A number of recent studies have provided support for the de novo gene birth processes for generation of orphan genes (Abrusan, 2013; Neme and Tautz, 2013; Arendsee et al., 2014; Neme and Tautz, 2014). Carnivus et al. (2012) detected the translation of hundreds of short species-specific open reading frames (ORFs) and argued that de novo origination of new genes could be more widespread than duplication (Carvunis et al., 2012).

A subset of de novo genes are those that arise via overprinting, i.e., a novel open reading frame originated within an existing gene through point mutations. While identification of de novo genes using homology-based method there is always a possibility to identify a highly diverged gene as a newly originated gene. Overprinted genes exclude such possibility as they retain the parental gene. Therefore, overprinted genes can serve as ‘true de novo’ genes and can be used to highlight the evolution of de novo genes. In this chapter, I have identified set of overprinted genes in A. thaliana. Figure 2.1 shows an A. thaliana orphan gene AT1G07485 that started coding within an older gene AT1G07490 using an alternative reading frame.
Figure 2.1: An example of a representative overprinted gene in *A. thaliana*. AT1G07485 is an orphan gene specific to *A. thaliana*. The gene has acquired an alternative reading frame within an older gene, AT1G07490.

For *S. bicolor*, we had access to transcriptome data of 40 accessions (unpublished data from Plant Genome Mapping Laboratory, University of Georgia). The germplasm of *S. bicolor* consists of five basic races: *bicolor, guinea, caudatum, kafir* and *durra* and their 10 hybrid races (*guinea-bicolor, caudatum-bicolor, kafir-bicolor, durra-bicolor, guinea-caudatum, guinea-kafir, guinea-durra, kafir-caudatum, durra-caudatum, kafir-durra*) (Wang et al., 2014). Figure 2.2 represents the areas of origin and possible migration routes for the *S. bicolor* races (Kimber et al., 2013). This 40 accessions transcriptomes data represents the five basic races and two hybrid races: *guinea-caudatum* and *durra-caudatum*. These accessions were randomly selected for transcriptome sequencing. I have used this RNA-seq data to identify: a) expression evidence of orphan gene sets in different accessions, b) sets of orphan genes displaying population-specific or accession-specific expression.
Figure 2.2: Areas of origin and development for the domesticated races of bicolor and their possible migration routes (Source: Kimber et al., 2013).
2.2 Materials and methods

2.2.1 Identification of orphan genes in *S. bicolor* and *C. sativa*

Peptide sequences of all representative gene models for *S. bicolor* (version 2.1) and all sequenced plant species were downloaded from the Phytozome database (Goodstein et al., 2012). *C. sativa* peptide sequences were provided by Isobel Parkin (Saskatoon Research Centre, Agriculture and Agri-Food Canada). A BLAST-based filtering approach was to identify orphan genes in both the species. BLASTP (Altschul et al., 1990) with an e-value cutoff of < 0.01 was used to search all predicted peptides of both the species against all the sequenced plant species (Phytozome database V9.1 http://phytozome.org/). The orphan candidates were filtered out and then BLAST searched against the NCBI nr, nt and est databases using BLASTP, TBLASTN and TBLASTN programs respectively (e-value < 0.01). The species name and family of all the orphan candidates were extracted from the NCBI taxonomy database using custom-made python scripts. Filtered candidates were further searched against the NCBI nr database using PSI-BLAST (Altschul et al., 1997). InterProScan 5 (Jones et al., 2014) was used to filter out orphan candidates with known domains. Figure 2.3 represents the search scheme for identification of orphan genes. The Irish Centre for High-End Computing (ICHEC) facility was used to perform all the BLAST searches and InterProScan.

2.2.2 Identification of evolutionary origins of orphan genes

The orphan genes that have originated by gene duplication events were identified by BLASTP against all the non-orphan genes (e-value < 0.01). For significant hits, HSPs were concatenated to get the query coverage. Hits with percentage coverage less than 10% were discarded.

An orphan gene can be considered to be *de novo* evolved if it shows significant intergenic or out-of-frame CDS hits with other sequenced plant species (but no significant sequence similarity with any pre-existing protein coding gene). Orphan genes containing intergenic or out-of-frame CDS hits were identified using BLASTN (e-value < 0.01). Orphan genes showing percentage coverage less than 10% were
discarded. Bedtools (Quinlan and Hall, 2010) were used to identify whether an orphan gene is showing a hit with a coding region or an intergenic region. A custom-made python script was used to identify the out-of-frame hits. An orphan gene was considered to be evolved by overprinting if its peptide sequence overlapped with the CDS (Coding Segment) of other genes (but not with UTRs or intronic regions). Bedtools were used to identify the overprinted genes models (Quinlan and Hall, 2010).

2.2.3 Expression evidence of sorghum-specific orphans in RNA-seq data of 40 accessions

Expression evidence of orphan genes identified in S. bicolor reference genome was obtained from RNA-seq data of 40 different accessions. RNA was extracted from three organs: leaf, stem, and flower. All the three samples were pooled together and short-read paired-end data was generated for all the accessions (read size 150 bp and insert-size 230 bp). Paired-end reads for each accession were mapped using Tophat 2.0.10 (Kim et al., 2013). Tophat mapping was done by providing the sorghum assembly and annotation version 2.1 and using the default mapping criteria (and --mate-inner-dist -70 bp as per our dataset). Cufflinks (Trapnell et al., 2010) was used to calculate the FPKM (fragments per kilobase of exon per million fragments mapped) which indicates the expression of each gene. Supplementary Figure 2.1 represents the FPKM distribution of 40 sorghum accessions. Genes with FPKM threshold >1 were considered to be expressed.

2.2.4 Identification of overprinted genes in A. thaliana

Bedtools (Quinlan and Hall, 2010) were used to identified all the overlapping gene models in the A. thaliana genome (TAIR v10). The overlapping gene models were categorized as: a) CDS-CDS, b) intron-CDS, c) UTR-CDS overlapping. As a first instance, I have only used CDS-CDS overlapping gene pairs. Only overlapping pairs that were translated in different reading frames were considered as overprinted genes based on BLASTTP hits.
2.2.5 Determining the gene age and rate of evolution of overprinted genes

Parental and newly arose overprinted genes from overprinted pairs were differentiated based on their gene age. BLASTP of overprinted genes against all the sequenced plant species was used to predict the gene age (e-value <1e-3). Age class was assigned to each overprinted gene pair where AC 0 represents the youngest age class and AC 9 as the oldest class. Age Class (AC): AC0- A. thaliana specific, AC1- A. lyrata, AC2-Brassicaceae, AC3- Brassicales-Malvales, AC4- Rosid, AC5- Eudicot, AC6-Angiosperm, AC7- Tracheophyte, AC8- Embryophyte, AC9- Viridiplantae.

Pair wise $dN/dS$ calculations between reciprocal best hits of A. thaliana and A. lyrata were adapted from Donoghue and Spillane (2013). Statistical tests were preformed as implemented in R package (r-project.org).
Figure 2.3: Schematic of search strategy for identification of orphan genes in a given genome.
2.3 Results

2.3.1 Identification of 3,761 Brassicaceae-specific orphan genes in C. sativa

C. sativa is an oil seed crop belonging to the Brassicaceae family. Despite its close evolutionary relationship with the model plant A. thaliana, C. sativa contains more than three times number of genes than A. thaliana due to its hexaploid genome structure (Hutcheon et al., 2010). The estimated genome size of C. sativa is 750 Mb and the number of chromosomes (n=20) is the highest among sequenced Brassicaceae genomes (Hutcheon et al., 2010). A total of 3,761 Brassicaceae-specific orphans were identified in C. sativa from 89,418 protein-coding gene models, which represents 4.2% of the C. sativa genome. Out of these, 1,656 orphans were identified to be C. sativa specific orphans (1.85% of the total number of predicted C. sativa genes).

Like orphan genes identified in other species, C. sativa orphan genes also displayed shorter sequence length, lower GC content and less introns, which are general sequence characteristics of orphan genes (Table 2.1) (Domazet-Loso and Tautz, 2003; Cai et al., 2006; Wilson et al., 2007; Toll-Riera et al., 2009; Donoghue et al., 2011; Varshney et al., 2011). There is no major bias observed in terms of distribution of orphan genes on the 20 chromosomes of C. sativa (Figure 2.4).

Table 2.1: Comparison of sequence features of orphans and non-orphans in the reference genome of C. sativa.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Orphans</th>
<th>Non-Orphans</th>
<th>P-value (U test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDS length (nucleotides)</td>
<td>278.57±3.513</td>
<td>1219.25±3.311</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Percent GC</td>
<td>41.89±0.126</td>
<td>44.95±0.011</td>
<td>1.00e-06</td>
</tr>
<tr>
<td>Percent GC1</td>
<td>44.23±0.195</td>
<td>50.21±0.016</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Percent GC2</td>
<td>39.54±0.188</td>
<td>40.61±0.019</td>
<td>2.00e-03</td>
</tr>
<tr>
<td>Percent GC3</td>
<td>41.88±0.181</td>
<td>44.03±0.023</td>
<td>5.00e-02</td>
</tr>
<tr>
<td>Number of introns</td>
<td>2.29±0.05</td>
<td>4.445±0.018</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>
Figure 2.4: Percentage of orphan genes on 20 chromosomes of *C. sativa*. *X-axis* represents the 20 chromosomes of *C. sativa*. *Y-axis* represents the % of orphan genes on each chromosome.

Using a BLAST-based sequence similarity approach, evolutionary origins of 2,169 of the 3,761 Brassicaceae-restricted orphan genes were identified. A total of 1,737 of these orphan genes have significant BLASTP hits with *C. sativa* non-orphan genes, which indicates that nearly half of *C. sativa* orphan genes have arisen by gene duplication of non-orphan genes. An additional 367 orphans were found to have a significant BLASTN hit at CDS level. A total of 354 (9.4%) orphans have significant BLASTN hits with other Brassicaceae species at either CDS or non-coding level. A small number of orphans (0.7%) displayed significant similarity to non-coding or out-of-frame CDS in non-Brassicaceae species. A total of 53 orphan genes were found to have arisen by overprinting (i.e. generating a new ORF from an existing ORF) of other non-orphan genes. For the orphan genes whose mode of origin could be identified, indicates that gene duplication from non-orphan genes was the
predominant (46.18%) mechanism by which orphan genes have arisen in the *C. sativa* genome.

**Table 2.2: Evolutionary origins of orphan genes in *C. sativa.***

<table>
<thead>
<tr>
<th>Mode of evolution</th>
<th>Number of orphan genes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplication</td>
<td>1,737 (46.18%)</td>
</tr>
<tr>
<td>Out-of-frame CDS hits in Brassicaceae (<em>C. sativa</em> specific)</td>
<td>210 (12.68%)</td>
</tr>
<tr>
<td>Non-coding region hits in Brassicaceae (<em>C. sativa</em> specific)</td>
<td>144 (8.69%)</td>
</tr>
<tr>
<td>Overlapping gene models</td>
<td>53 (1.41%)</td>
</tr>
<tr>
<td>Out-of-frame CDS hits in non-Brassicaceae</td>
<td>16 (0.43%)</td>
</tr>
<tr>
<td>Non-coding region hits in non-Brassicaceae</td>
<td>9 (0.24%)</td>
</tr>
</tbody>
</table>
2.3.2 Identification of 2,499 species-specific orphan genes in *S. bicolor*

*S. bicolor* represents the first crop genome of African origin to be sequenced, a diploid grass crop species (2n=20) with a genome size of ~730 Mb (Paterson et al., 2009). *S. bicolor* is a widely grown cereal crop, mainly in Asia, Africa and Central America. *S. bicolor* belongs to the grass family Poaceae and is closely related to *Zea mays* and sugarcane (Chapter 1, Figure 1.4) (Paterson et al., 2009). The sorghum genus has evolved across a wide range of environments in Africa and exhibits a range of phenotypic diversity and numerous resistances to abiotic and biotic stress conditions (Kimber et al., 2013).

I have identified only species-specific orphan genes in *S. bicolor* as these genes will be further analyzed using recently sequenced outgroup species at a later date (unpublished data) for evidence of gene model retention. Orphan genes specific to *S. bicolor* (excluding outgroups) will be further screened for functionality to identify *S. bicolor* specific adaptations. Using the BLAST-based stepwise filtering approach (Figure 2.3), I have identified 2,499 *S. bicolor*-specific orphan genes from 33,032 protein-coding gene models defined in the sorghum genome (annotation version 2.1). Hence, the sorghum-specific orphan genes represent 7.6% of the total number of protein-coding genes.

As in *C. sativa*, orphan genes identified in *S. bicolor* genome displayed shorter sequence length, atypical GC content, and also a lower number of introns compared to non-orphan genes (Table 2.3). No major bias towards residing on specific chromosomes was observed in terms of distribution of the sorghum orphan genes across the 10 chromosomes of sorghum genome (Figure 2.5; χ²-test, p= 0.9514).
Table 2.3: Comparison of sequence features of orphans and non-orphan genes in the reference genome of *Sorghum bicolor*.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th></th>
<th>P-value (U test)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orphans</td>
<td>Non-Orphans</td>
<td></td>
</tr>
<tr>
<td>CDS length</td>
<td>335.27±3.108</td>
<td>1246.41±5.174</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Percent GC</td>
<td>56.17±0.194</td>
<td>56.99±0.056</td>
<td>4.00e-03</td>
</tr>
<tr>
<td>Percent GC1</td>
<td>56.74±0.220</td>
<td>58.68±0.044</td>
<td>1.30e-10</td>
</tr>
<tr>
<td>Percent GC2</td>
<td>54.91±0.226</td>
<td>46.52±0.045</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Percent GC3</td>
<td>56.87±0.216</td>
<td>65.77±0.108</td>
<td>&lt; 2.2e-16</td>
</tr>
<tr>
<td>Number of introns</td>
<td>0.63±0.017</td>
<td>3.78±0.028</td>
<td>&lt; 2.2e-16</td>
</tr>
</tbody>
</table>

Figure 2.5: Orphan genes (as percentage of genes on each chromosome) located on each of the 10 chromosomes of sorghum (\(\chi^2\)-test, p= 0.9514, df=10). *X-axis represents the 10 chromosomes of S. bicolor. Y-axis represents the % of orphan genes on each chromosome.*

48
Using BLAST-based sequence similarity approach, the evolutionary origin of the *S. bicolor* orphan genes was investigated. Two modes of evolution of orphan genes were considered: (a) gene duplication and (b) *de novo* gene origination (including via overprinting). This analysis indicated that 297 orphans have originated by duplication from *S. bicolor* non-orphan genes (Table 2.4). Amongst these 297 orphans, 183 orphans showed significant hits with non-orphan genes at peptide (protein sequence) level, while 114 orphans showed significant CDS (DNA sequence) hits with the non-orphan genes.

To identify orphan genes of *de novo* origin, evidence of orphan gene evolution from non-coding sequences was identified (from intergenic regions and from frame-shift ORF generation). Based on BLASTN search, a total of 196 orphans were found to have arisen from intergenic regions of other sequenced plant species (Phytozome database V9.1 http://phytozome.org/). Furthermore, 70 orphan genes had significant out-of-frame hits with the other sequenced plant species at CDS level. None of the 2,499 orphan genes in the sorghum genome have originated by overprinting of other orphan or non-orphan genes.

**Table 2.4: Modes of evolutionary origin of orphan genes in *S. bicolor*.**

<table>
<thead>
<tr>
<th>Mode of evolution</th>
<th>Number of orphans (% of total orphans)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplication</td>
<td>297 (11.88%)</td>
</tr>
<tr>
<td><em>De novo</em> (from intergenic regions)</td>
<td>196 (7.84%)</td>
</tr>
<tr>
<td><em>De novo</em> (frameshift ORF generation)</td>
<td>70 (2.8%)</td>
</tr>
<tr>
<td><em>De novo</em> (via overprinting)</td>
<td>0</td>
</tr>
</tbody>
</table>

To identify orphan genes supported by gene expression evidence, RNA-seq data for 40 *S. bicolor* accessions was analyzed. A total of 1,871 orphan genes were expressed in at least one of the 40 accessions, which represents 74.9% of the total 2,499 orphan genes identified. A homogeneous distribution of expressed orphans is observed across all the accessions (Figure 2.6; $\chi^2$-test, p= 1, df=39). A total of 52 (2.1%) orphan genes were expressed in all of the sorghum accessions and represents the ‘core set’ of expressed orphan genes in sorghum.
None of the *S. bicolor* orphan genes showed population-specific expression (i.e. specific to populations: *bicolor*, *guinea*, *caudatum*, *kafir*, *durra*, *guinea-caudatum* and *durra-caudatum*). However, a set 420 orphan genes showed accession-specific expression i.e., expressed in only one accession.

![Graph showing percentage of orphan genes expressed in each accession](image)

**Figure 2.6:** Percentage of orphan genes expressed in each of the *S. bicolor* accessions. *X-axis* represents the 40 accessions of *S. bicolor*, % of expressed orphan genes in each accession are shown on *Y-axis.*
2.3.3 Accelerated evolutionary rate of De novo originated overprinted genes in *A. thaliana*

Further I have identified overprinted genes in model plant species *A. thaliana*. A total of 989 overlapping gene pairs were identified from 27,416 gene models in *A. thaliana* that represents 3.6% of the total protein-coding genes in *A. thaliana* genome. Out of these 663 pairs (67%) were resulted from UTR-CDS overlap, 220 (22.2%) from intron-CDS overlap and 106 (10.7%) from CDS-CDS overlap (Figure 2.7). To develop the first prototype of the pipeline for identification and evolution of overprinted genes, I have firstly used CDS-CDS overlapping genes.

As per the definition of overprinted genes, these are the overlapping genes that start coding within an existing gene by making use of an alternative reading frame (Keese and Gibbs, 1992). To identify overprinted genes from CDS-CDS overlapping pairs, a BLASTP analysis was conducted to extract overlapping gene pairs that were translating in different reading frames. Overlapping pairs that were translating in the same reading frame were discarded. Out of 106 CDS-CDS overlapping pairs, 80 gene pairs were identified to be translating in different reading frames and were considered as overprinted genes. Parental and newly arisen overprinted genes were classified from 80 overprinted gene pairs based on their gene age (Supplementary Table 2.1). BLASTP analysis was used to predict the gene age of overprinted genes (e-value <1e-3). These preliminary results show that overprinted genes in *A. thaliana* are fast evolving when their evolutionary rate is compared with all *A. thaliana* protein coding genes (U-test, P=0.8e-3) (Table 2.5).

**Table 2.5**: Evolutionary rate of *A. thaliana* overprinted genes in comparison to the whole genome as background.

<table>
<thead>
<tr>
<th></th>
<th>Mean $\omega (dN/dS)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overprinted genes</td>
<td>0.4000±0.0715</td>
</tr>
<tr>
<td>Background</td>
<td>0.2437±0.0022</td>
</tr>
</tbody>
</table>
Figure 2.7: Number of overlapping gene pairs identified in the *A. thaliana* reference genome. *X-axis* represents the overlap category (CDS-CDS, Intron-CDS, UTR-CDS). Number of overlapping gene models in each category are shown on *Y-axis.*
2.4 Discussion

Each species contains a certain proportion of genes that are not present in any other species and are unique to the species, or even to certain populations within it. The percentage estimates of these orphan genes in various species ranges from <1–71% with 5–15% being fairly typical (Arendsee et al., 2014). I have identified 1.85% (1,656 genes) and 7.6% (2,499 genes) orphan genes in the genomes of two crop species C. sativa and S. bicolor respectively. Orphan genes identified in other plant species includes 3.4% (958 genes) of those in A. thaliana (Donoghue et al., 2011), 0.35% (169 genes) of C. cajan (Varshney et al., 2012), and 3.2% (1,926 genes) of O. sativa (Guo et al., 2007). A higher percentage (7.6%) of orphan genes in S. bicolor (though not unusual) could be due to recent expansion of a few gene families in S. bicolor (unpublished data, Plant Genome Mapping Laboratory, University of Georgia). Orphan genes are a subset of taxonomically restricted (also called lineage-specific) genes that are specific to a particular taxon (e.g., malvid-specific or Brassicaceae-specific genes) (Arendsee et al., 2014). In addition to species-specific genes, an additional 2,105 Brassicaceae-restricted orphan genes were identified in C. sativa.

Orphan genes identified by previous studies in different species displayed common sequence characteristics. Shorter sequence length, faster evolution, atypical GC content and lower number of introns is considered as general sequence characteristics of orphan genes (Domazet-Loso and Tautz, 2003; Guo et al., 2007; Wilson et al., 2007; Donoghue et al., 2011). Orphan genes identified in S. bicolor and C. sativa displayed similar trend of sequence characteristics, despite of belonging to two different groups of angiosperms (or flowering plants), monocots and dicots respectively. Monocots branched off from dicots 150 mya (Chaw et al., 2004). Newly arisen genes are generally expected to have shorter length and lack of introns due to improbability of a long ORF and the complexity of introns splicing signals (Knowles and McLysaght, 2009). UTR (untranslated region) introns are considered to be more easily acquired compared to coding region introns due to lower constraints (Hong et al., 2006).
Systematic identification of evolutionary origins of orphan genes in *A. thaliana* has revealed that a quarter of *A. thaliana* orphan genes were originated from duplication of non-orphan genes and almost half of orphan genes originated *de novo* (Donoghue et al., 2011). However over 20% of *C. sativa* orphan genes were identified to be originated through *de novo* evolution, so duplication can be concluded to be the major mechanism of origination of *C. sativa* (46%) orphan genes. The high number of duplicated orphan genes in *C. sativa* could be due to whole-genome triplication of *C. sativa* genome. In *S. bicolor*, the evolutionary origin of only 22% orphan genes was identified, where both *de novo* (10.6%) and duplication (11.9%) mechanisms equally participated. The lack of evidence of mechanism of origin for *S. bicolor* orphans could be due to rapid evolution of orphan genes. It is also possible that some orphan genes have diverged beyond the point at which their evolutionary origins can be found by BLAST. I could not identify any *S. bicolor* orphan genes to evolving via overprinting due to the absence of annotated overlapping gene models in *S. bicolor* annotation (latest annotation version 2.1). Identification of overprinted orphan genes will be revisited in an improved version of *S. bicolor* annotation.

Similar to *A. thaliana* orphan genes (Donoghue et al., 2011), overprinted genes identified in *A. thaliana* have also demonstrated rapid evolution. These results are consistence with the previous findings that newly arisen genes evolve very rapidly after their origin (Guo et al., 2007; Sabath et al., 2012; Arendsee et al., 2014). The pipeline for overprinted genes will be further extended to introns-CDS and CDS-UTR overlapping pairs and transcription and translation evidence of overprinted genes will be identified.

In the next chapter, I have identified a number of chimeric open reading frames (ORFs) in *Cajanus cajan*. The novel chimeric ORFs produced as a result of mitochondrial genome rearrangements is considered to be the main cause of cytoplasmic male-sterility (CMS) in plants. The chimeric ORFs identified in male-sterile lines of *Cajanus cajan* mitochondrial genome represents the most promising candidates for cytoplasmic male-sterility related rearrangements.
2.5 Conclusions and future work

*C. sativa* and *S. bicolor* are important agronomic crops. Along with their broad distribution, both species have the potential to adapt to future climate conditions, particularly increasing drought. Identification of genes that are unique to such species (orphan gens) can provide an important set of genes that might have played a role in acquiring adaptation to different environmental conditions and other unique characteristics to these species. Therefore, the research on orphan genes in agronomically important crops not only implicates meaningful evolutionary fundamentals of species but also can assist plant breeding. Orphan genes identified in *S. bicolor* will be further analyzed in recently sequenced outgroup species for identification of gene models retention. Orphan genes utterly unique to *S. bicolor* will be further screened for functionality to identify *S. bicolor* specific adaptations.

In this chapter, I have also presented the first prototype of the pipeline for identification of overprinted genes in *A. thaliana* and demonstrated the rapid evolution of these genes. This work will be further extended to introns-CDS and CDS-UTR overlapping pairs. Young and old overprinted genes will be categorized for studying their evolutionary rates and population genetics analysis in *A. thaliana* population data. The transcription and translation evidence of these genes will be identified. Any loss-of-function candidate overprinted genes will be identified in *A. thaliana* population data.
Chapter 3. Cytoplasmic male sterility associated chimeric ORFs identified by mitochondrial genome sequencing of four Cajanus genotypes


Author contributions:
RT and RKS performed experiments, analyzed data and contributed to writing the manuscript; JD, TS, WC, GF, AJA and CS contributed to analysis data and interpretation of result; YLX, CDT performed experiments and analyzed data, and RKV conceived and guided experiments, contributed to analysis of data, interpreted results, contributed to writing and the finalized manuscript.

* I have performed all the bioinformatics analysis for this paper, except section 3.2.5 (comparison of gene order with other plant species).
Abstract

The hybrid pigeonpea (*Cajanus cajan*) breeding technology based on cytoplasmic male sterility (CMS) is currently unique amongst legumes and displays major potential for yield increase. CMS is defined as a condition in which a plant is unable to produce functional pollen grains. The novel chimeric open reading frames (ORFs) produced as a result of mitochondrial genome rearrangements is considered to be the main cause of CMS. To identify these CMS-related ORFs in pigeonpea, we sequenced the mitochondrial genomes of three *C. cajan* lines (the male-sterile line ICPA 2039, the maintainer line ICPB 2039, the hybrid line ICPH 2433) and of the wild relative (*C. cajanifolius* ICPW 29). A single, circular-mapping molecule of length 545.7 Kb was assembled and annotated for the ICPA 2039 line. Sequence annotation predicted 51 genes, including 34 protein-coding and 17 RNA genes. Comparison of the mitochondrial genomes from different *Cajanus* genotypes identified 31 ORFs, which differ between lines within which CMS is present or absent. Among these chimeric ORFs, 13 were identified by comparison of the related male-sterile and maintainer lines. These ORFs display features which are known to trigger CMS in other plant species and represent the most promising candidates for CMS-related mitochondrial rearrangements in pigeonpea.

**Key words:** mitochondria, pigeonpea, next generation sequencing, cytoplasmic male sterility, open reading frames
3.1 Introduction

Angiosperm mitochondrial genomes are unique in eukaryotes because of their high rates of rearrangement, sequence duplication, ongoing gene loss, and frequent incorporation of foreign DNA (Adams and Palmer, 2003; Kubo and Newton, 2008; Bock, 2010). Land plant mitochondrial genomes vary in size from 105 Kb (Terasawa et al., 2007) to more than 11,000 Kb (Sloan et al., 2012). Hence, the smallest land plant mitochondrial genome (*Physcomitrella patens*, 105 Kb) is still ~11 times larger than the human mitochondrial genome (Anderson et al., 1981) (~16 Kb). Several studies have reported the presence of subgenomic circles in mitochondrial genomes that have arisen from recombination events (Nair, 1993; Backert et al., 1996). While such recombination events in plant mitochondria increase the complexity of their genome structures, recombination has also been proposed to maintain genomic stability and may also provide a mechanism to increase genetic variation in the absence of sexual reproduction (Mackenzie and McIntosh, 1999; Mach, 2011).

Rearrangements in mitochondrial genomes are of considerable biotechnological interest as they can cause cytoplasmic male sterility (CMS), which is a valuable tool for plant breeding programmes. Male-sterility is caused by the failure of a plant to produce functional pollen grains. CMS is a maternally inherited trait and is mainly controlled by the mitochondrial genome. CMS is often found to be caused by chimeric mitochondrial open reading frames (ORFs) that are produced as a result of mitochondrial genome rearrangements (Zabala et al., 1997; Schnable and Wise, 1998; Hedgcoth et al., 2002; Wang et al., 2006). In many cases of CMS, male fertility can be restored by the introduction of nuclear genes known as restorer-of-fertility (*Rf*) genes.

Although plant breeders have used CMS technology for producing F1 hybrids for enhancing crop productivity in numerous cereal and vegetable crops, the development of F1 hybrids has not been possible in legumes because of their high levels of self-pollination. In pigeonpea, however, a moderate level of insect-mediated out-crossing exists that could be used to develop a stable CMS system. In 2005, Saxena *et al.* (Saxena et al., 2005) derived a stable CMS system, ICPA 2039A, from an interspecific hybrid of cultivated pigeonpea (*Cajanus cajan*) and a wild relative (*C.
cajanifolius) (Supplementary Figure 3.1). Previous CMS systems have been attempted in pigeonpea (Ariyanayagam et al., 1995; Saxena and Kumar, 2003), but have been unsuccessful, mostly as a result of instability in the expression of male-sterility and fertility restoration (Saxena et al., 2005). The development and utilization of stable male-sterile lines from different cytoplasmic backgrounds is a key factor to the diversification of pigeonpea hybrid parental lines. Indeed, male-sterility systems in many crops do not allow generation of completely male-sterile progenies, drastically limiting use of male-sterile lines in F1 hybrid seed production (Li and Yuan, 2004). To accelerate hybrid pigeonpea breeding for yield and quality, understanding the molecular basis of male-sterility is critically important. Specifically, the identification of CMS-associated genetic polymorphisms is a key prerequisite for rational development of new and improved CMS systems for the production of superior F1 hybrids. Next-generation sequencing (NGS) has provided opportunities to gain the genetic information in a much faster and cost effective manner. NGS of mitochondrial genomes and analysis of genetic variations across the genomes of male-sterile, maintainer, and wild relative species will facilitate the identification of genetic features related to male sterility.

This study reports the generation and analysis of mitochondrial genome sequences of four Cajanus genotypes: the male-sterile line ICPA 2039, the maintainer line ICPB 2039, the hybrid line ICPH 2433, and the wild relative ICPW 29 (C. cajanifolius). A high-quality pigeonpea mitochondrial genome assembly has been developed for ICPA 2039. This study provides the first comparative study of legume mitochondrial genome sequences and identifies several re-arrangements and no-coverage regions (large regions >1,000 bp; with zero coverage), as well as chimeric ORFs associated with CMS in pigeonpea.
3.2 Materials and methods

3.2.1 Plant material and mitochondrial DNA isolation

Cajanus lines ICPA 2039, ICPB 2039, ICPH 2433 and ICPW 29 were used as the source of mitochondrial DNA (mtDNA). mtDNA was isolated from two-week-old etiolated seedlings and was purified before sequencing (Scotti et al., 2001).

3.2.2 Sequencing and assembly

Mitochondrial genomes of four pigeonpea lines were pyrosequenced with the Roche/454 FLX sequencing platform following whole-genome amplification (WGA). WGA kit GenomePlex from Sigma (Sigma-aldrich, St. Louis, USA) was used in this study. Twenty nanograms of DNA template were used for WGA according to the protocol from manufactures. In summary, the WGA process was divided into fragmentation, library generation, and PCR amplification. The first two steps, fragmentation and library generation (3 kb of insert size), were carried out without interruption, to avoid the DNA degradation. Further to amplify higher amount of DNA, the GenomePlex reaction was allowed to proceed for 4 h. De novo genome assembly of reference genome (ICPA 2039) was performed using Newbler, Celera and CLC bio software programs. All the usable reads were aligned onto the contig sequences and aligned paired-end sequences (PEs) were obtained. We then calculated the amount of shared PE relationships between each pair of contigs, weighted the rates of consistent and conflicting PEs, and then constructed the scaffolds step-by-step, beginning with the shortest insert-sized PEs, to long insert-sized PEs. Assemblies generated by Newbler assembler were considered most robust in terms of length of the scaffolds and genome coverage and were used for further analysis. Gaps within the assembly were identified using contig-graph information. The Perl script, parse_link.pl was used to identify and close the gaps insilico (Nagarajan et al., 2010) (http://www.cbcb.umd.edu/finishing/finishing-v1.tar.gz). Remaining gaps were filled by Sanger sequencing. Graphs were generated for a preliminary view of the assembly (Figure 3.1) in an effort to check the order and orientation of mitochondrial scaffolds in the genome. Scaffolds that were not connected to other scaffolds in graph and showed low coverage were suspected to be part of chloroplast genome. BLASTN
searches were performed for these scaffolds against NCBI database to validate these scaffolds are contamination from chloroplast genome. Assembly graphs were used as a guide to connect the scaffolds. Primers were designed from the ends of the scaffolds that showed connections with other scaffolds in assembly graphs. The orientation of each scaffold within the assembly was confirmed by Sanger sequencing.

Figure 3.1: A scheme showing linking scaffold with the help of graph in a preliminary view of the assembly. Assembly graphs were used as a guide to connect the scaffolds. Each box represent a scaffold, ‘||’ represent the 3’ end, and ‘|>’ represent the 5’ end of each scaffold. Number on each scaffold represents the scaffold number and size of the scaffold. The thick black lines indicate that the scaffolds are attached in correct orientation and spotted lines indicate that the scaffolds are attached in reverse orientation in the assembly. Numbers on these lines represents the sequence coverage. The orientation of each scaffold was confirmed through Sanger sequencing.
3.2.3 Gene prediction and annotation

Protein-coding and RNA genes were predicted by performing BLASTX and BLASTN searches, respectively, against a database of protein-coding, tRNA and rRNA genes complied from all previously sequenced seed plant mitochondrial genomes (Alverson et al., 2010). tRNAscan-SE (Lowe and Eddy, 1997) was used to corroborate the tRNA boundaries identified by BLASTN. A BLAST score of e-value <1e-3 and percent identity threshold of >70% were initially used for filtering BLAST outputs. Gene boundaries were extended or trimmed to the positions of the start and stop codons manually using Artemis 12.0 (Rutherford et al., 2000). Annotation data were written to a Sequin-formatted table file with a set of Perl and CGI scripts.

3.2.4 Gene order comparison

In order to identify colinearity between the pigeonpea mitochondrial genome and other angiosperms, we used BLAT (Standalone BLAT v. 34) with an identity cut-off of ≥0.9 and coverage ≥0.5 to compare the gene order of our Cajanus genomes in a pairwise fashion to that of the following 11 angiosperms: Vigna radiata (Alverson et al., 2011b), Triticum aestivum (Liu et al., 2011), Oryza sativa (Notsu et al., 2002), Zea mays (Clifton et al., 2004), Arabidopsis thaliana (Unseld et al., 1997), Beta Vulgaris (Kubo et al., 2000), Citrullus lanatus (Alverson et al., 2010), Cucurbita pepo (Alverson et al., 2010), Nicotina tabacum (Sugiyama et al., 2005), Vitis vinifera (Goremykin et al., 2009), Cucumis sativus (Alverson et al., 2011a).

3.2.5 Genome alignment and representation

The scaffolds of the three other mitochondrial genomes (ICPB 2039, ICPH 2433 and ICPW 29) were aligned with the finished assembly of ICPA 2039 genome using BLASTN. Circular genome representations of ICPA 2039 genome were generated using OGDRAW (Lohse et al., 2007). Three different maps were generated to represent the alignment of ICPA 2039 genome with the three other genomes. Figures were scaled down to integrate all the four maps in a single map.
3.2.6 Identification of rearrangements and no-coverage regions

Comparative assemblies of all three genomes were generated using GS Reference Mapper 2.5. Raw reads of each mt genome were aligned to that of ICPA 2039 genome assembly in order to detect any sequence-level differences between them. Rearrangements with >60% frequency was considered for further analysis. No-coverage regions were extracted using a custom Perl script, which checks the coverage of every base in the assembly and groups the consecutive positions where coverage is very low. Regions larger than 1 kb and with approximately zero coverage were considered as no-coverage regions.

3.2.7 Chimeric ORFs

Sequences for Open Reading Frames (ORFs) >100 codons in the vicinity of rearrangements or within no-coverage regions were collected using Artemis 12.0 (Rutherford et al., 2000). ORFs coding for known mitochondrial genes were excluded from the analysis. Further, these ORFs were blasted against the ICPA 2039 genome itself in order to check whether these ORFs carry part of other genes or ORFs. First hit of the blast match were left, as that will be the original location of these ORFs. All the other hits showing identity ≥95% and sequence coverage ≥16 bp were considered. Further, these ORFs were checked in terms of their closeness to any predicted gene. Potential transmembrane helices were predicted with TMHMM 2.0 (Krogh et al., 2001). A scoring criteria from 0-4 was assigned to each ORF, one for presence of parts of other genes, one for proximity of any predicted genes, one for presence of hydrophobic domains and one additional score for carrying parts of atp genes. ORFs showing score ≥3 were considered as the potential chimeric ORFs.
3.3 Results and discussion

Pigeonpea is an important legume crop for resource-poor smallholder farmers in marginal environments. Unfortunately, the productivity of this legume staple crop has stagnated at ca. 750 Kg/ha due to its exposure to biotic and abiotic stresses. Pigeonpea recently became the first legume to have F₁ hybrids released based on a CMS system (Stokstad, 2007). The initial pigeonpea F₁ hybrids (e.g. ICPH 2671) showed >30% yield advantages over the best pure line varieties in the same geographic regions. Such advances clearly indicate that pigeonpea F₁ hybrid technology has the potential to break the current yield plateau. For successful and sustainable pigeonpea hybrid production and extension, the following are critical factors: (i) diversification of parental lines and CMS sources, (ii) improvement of parental lines for tolerance to biotic and abiotic stresses, and (iii) maintaining the purity of hybrid seeds. In this context, improvement of parental lines is underway through conventional and molecular breeding approaches that are being accelerated by the availability of a sequenced pigeonpea genome (Varshney et al., 2011). Simple sequence repeat (SSR) markers-based F₁ hybrid purity testing has also been initiated for ensuring purity of hybrid seeds (Saxena et al., 2010b; Bohra et al., 2011). However, major challenges remain in relation to the need for diversification of CMS sources in the pigeonpea gene pool. Although seven cytoplasmic sources are available (C. sericeus, C. scarabaeoides, C. volubilis, C. cajanifolius, C. cajan, C. lineatus, and C. platycarpus), only C. cajanifolius has currently been commercially exploited. The other six sources have not been able to be used commercially, because they express the CMS trait at an adequate level. To understand the factors conditioning the efficiency differences between sources of CMS, it will first be necessary to understand the molecular basis of CMS in pigeonpea.

3.3.1 Sequencing and assembly of four mitochondrial genomes of pigeonpea

We used Roche/454 FLX technology, targeted Sanger sequencing, and computational approaches to produce one complete and three draft assemblies for the mitochondrial genomes of four Cajanus lines. Northern blot-based screening and shotgun sequencing have been the conventional approaches used to identify CMS-associated chimeric ORFs in different plant species, including maize (Allen et al., 2007), sugar
beet (Satoh et al., 2004), rice (Tian et al., 2006; Liu et al., 2007), wheat (Liu et al., 2011), and brassica (Chen et al., 2011). Recently, Bentolila and Stefanov (Bentolila and Stefanov, 2012) successfully identified a candidate for the wild abortive CMS-encoding gene by pyrosequencing of two rice mitochondrial genomes using Roche/454 sequencing technology. However, Northern screening approaches used for the same rice genomes failed to identify any potential CMS candidates in wild abortive rice CMS lines (Bentolila and Stefanov, 2012). Therefore, we used Roche/454 sequencing technology and a de novo assembly approach to sequence the mitochondrial genome of Cajanus species. Due to the unknown genome architecture of the pigeonpea mitochondrial genome, we did not rely entirely on in silico approaches, but also validated proposed interscaffold connections by Sanger sequencing. To identify regions which varied in a manner correlated with CMS, the four mitochondrial genomes were then compared using the ICPA 2039 assembly as a reference.

Roche/454 sequencing of four genomes from purified mtDNA generated totals of 38.8 Mb, 15.6 Mb, 37.1 Mb and 23.8 Mb of paired-end data for ICPA 2039, ICPB 2039, ICPH 2433 and ICPW 29 respectively. The sequencing reads were assembled into scaffolds using three different de novo assembly programs – Newbler, CLCBio and Celera. Assemblies generated by the Newbler assembler were considered as the best assemblies (Table 3.1), with scaffold N50 values of 169.6 Kb, 1.2 Kb, 169.9 Kb and 108.1 Kb for ICPA 2039, ICPB 2039, ICPH 2433 and ICPW 29 respectively. Additional sequence data were generated for the comparatively under-sequenced genotypes of ICPB 2039 (15.1 Mb) and ICPW 29 (23.2 Mb). The sequencing data for these was reassembled and N50 of two assemblies thus improved from 1.2 Kb to 12.8 Kb for ICPB 2039 and from 108.2 Kb to 159.2 Kb for ICPW 29. In summary, mean scaffold lengths of 22.4 Kb, 8.8 Kb, 6.3 Kb and 37.3 Kb were achieved for the pigeonpea mitochondrial genomes of lines ICPA 2039, ICPB 2039, ICPH 2433 and ICPW 29 respectively. Analysis of sequence data for GC content indicated similar GC content distribution in all of the four genomes (Supplementary Figure 3.2).
<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Number of sequence reads (length) generated</th>
<th>Newbler</th>
<th>Celera</th>
<th>CLC Bio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of Scaffolds</td>
<td>Bases in Scaffolds (bp)</td>
<td>N50 Scaffold Size (bp)</td>
<td>Number of Scaffolds</td>
</tr>
<tr>
<td>ICPW 29</td>
<td>164,071 (53.4Mb)</td>
<td>18 (4)#</td>
<td>672,137 (575,487)##</td>
<td>159,243</td>
</tr>
<tr>
<td>ICPA 2039</td>
<td>121,170 (38.8Mb)</td>
<td>30 (7)#</td>
<td>672,918 (532,372)##</td>
<td>169,595</td>
</tr>
<tr>
<td>ICPB 2039</td>
<td>117,163 (36.4Mb)</td>
<td>52 (34)#</td>
<td>459,802 (335,926)##</td>
<td>12,823</td>
</tr>
<tr>
<td>ICPH 2433</td>
<td>116,021 (37.1Mb)</td>
<td>108 (9)#</td>
<td>681,810 (539,865)##</td>
<td>169,903</td>
</tr>
</tbody>
</table>

#Number of scaffolds from mitochondrial genome
##Base in scaffolds from mitochondrial genome
3.3.2 Finishing of reference mitochondrial genome ICPA 2039

The scaffolds of ICPA 2039 were further refined by removing contamination from nuclear or chloroplast genomes, and by closing gap regions within these scaffolds using Sanger sequencing. A preliminary view of the ICPA 2039 assembly was generated to check the connections between the scaffolds and for removal of contaminants (Figure 3.1). Scaffolds from the mitochondrial genome assembly were selected based on their coverage and links with the other scaffolds as shown in Figure 3.1. Selected scaffolds were subjected to BLASTN analysis against the NCBI database. Out of 30 selected scaffolds of ICPA 2039, with a total length of 672,918 bp, seven were confirmed to be mitochondrial in origin, representing 532,372 bp or 79% of the total sequence data (Table 3.1). Seven scaffolds were homologous to cpDNA, representing 78,461 bp (11.6% of the total sequence). Eleven further scaffolds matched nuclear DNA representing 46,228 bp (6.9% of the sequence). Two scaffolds representing 8,268 bp (1.2% of the sequence data) matched the sequence of the plasmid DNA. The remaining three scaffolds representing 8,309 bp (1.2%) of sequence data did not show any match in the NCBI database and may represent sequences unique to pigeonpea. The seven scaffolds that matched other plant mitochondrial DNA were targeted for further analysis. A total of 38 gaps (26,830 bp) were observed in the seven mitochondrial scaffolds, represented by Ns in the assemblies. Using the parse_link.pl script (http://www.cbcb.umd.edu/finishing/finishing-v1.tar.gz) from the finishing toolbox (see Materials and methods), 47 contigs from contigs that had not previously been assembled into scaffolds were introduced to fill 38 gaps inside the seven scaffolds. Two gaps (of 65 bp each), which the script was unable to fill in silico, were closed using Sanger sequencing technology. Subsequently, assembly graphs were used as a guide to connect the scaffolds. To confirm the order and orientation of each scaffold within the assembly, primer pairs were designed from the ends of each scaffold based on their connections with other scaffolds in assembly graphs. A set of 24 primer pairs (Supplementary Table 3.1) was used to generate amplicons and sequence data was generated for these using Sanger sequencing technology. In this way, a high quality, circular-mapping mitochondrial genome of 545,742 bp in total length, with ~23-fold coverage, was assembled for ICPA 2039. This master circular molecule contains a large recombinationally active repeat of size 4,951 bp which is extending from
positions 531,745 to 536,696 bp. Recombinationally active large repeats are a very common feature of plant mitochondrial genomes (Palmer and Shields, 1984; Davila et al., 2011).
3.3.3 Gene content

Within the mitochondrial genome of ICPA 2039, we identified 34 protein-coding, 14 tRNA and 3 rRNA genes (Supplementary Table 3.2), for a total of 29,346 bp of protein exons; 31,018 bp of intronic sequence; 5,255 bp of rRNA genes; and 1,477 bp of tRNA genes (Table 3.2). We did not find a copy of cox2, confirming proposals that this gene has lost in the legume lineages (Adams et al., 1999; Alverson et al., 2011b). In contrast, we found ICPA 2039 to contain two identical copies of the cox3 gene. Multiple copies of tRNAs for cysteine, lysine and methionine were observed to be present in the mitochondrial genome of ICPA 2039. The tRNA genes, carrying methionine are also highly similar to those of the plastid as is likely derived from the cpDNA, as is that of tryptophan (Supplementary Table 3.2).

To annotate mtDNA-encoded genes in the other sequenced lines, their scaffolds were compared with ICPA 2039. Six scaffolds derived from the ICPH 2433 hybrid contained 32 protein-coding genes and 12 tRNA genes between them (Supplementary Table 3.3). Four mitochondrial scaffolds of ICPW 29 covered 33 protein coding and 14 tRNA genes (Supplementary Table 3.4). As in ICPA 2039, multiple copies of cysteine, lysine and methionine-tRNA genes were present in the scaffolds of ICPH 2433 and ICPW 29. Due to low sequence coverage and small scaffold size, the 17 mitochondrial scaffolds of ICPB 2039 line only covered 15 protein and 11 tRNA genes between them (Supplementary Table 3.5).

Table 3.2: Genome coverage by coding features in ICPA 2039 mitochondrial genome assembly.

<table>
<thead>
<tr>
<th>Class</th>
<th>Feature</th>
<th>ICPA 2039* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>Total Size</td>
<td>545,742 bp</td>
</tr>
<tr>
<td>Coding</td>
<td>Protein exons</td>
<td>29,346 bp (5.4)</td>
</tr>
<tr>
<td>Coding</td>
<td>Introns</td>
<td>31,018 bp (5.6)</td>
</tr>
<tr>
<td>Coding</td>
<td>rRNA</td>
<td>5,255 bp (0.9)</td>
</tr>
<tr>
<td>Coding</td>
<td>tRNA</td>
<td>1,477 bp (0.2)</td>
</tr>
<tr>
<td>Non-coding</td>
<td>Mitochondria-like</td>
<td>220,747 bp (40.5)</td>
</tr>
<tr>
<td>Non-coding</td>
<td>Nuclear-like</td>
<td>40,330 bp (7.4)</td>
</tr>
</tbody>
</table>

*Figure in parentheses represents the percentage of total size*
3.3.4 Structural features of the pigeonpea mitochondrial genome compared to other plant species

The assembled *Cajanus* mitochondrial genome was compared with the mitochondrial genomes of 11 other land plant species. The species for comparison include one legume- *Vigna radiata* (Alverson et al., 2011b), three cereals - *Triticum aestivum* (Liu et al., 2011), *Oryza sativa* (Notsu et al., 2002), *Zea mays* (Clifton et al., 2004) and seven other eudicots, including *Arabidopsis thaliana* (Unseld et al., 1997), *Beta Vulgaris* (Kubo et al., 2000), *Citrullus lanatus* (Alverson et al., 2010), *Cucurbita pepo* (Alverson et al., 2010), *Nicotina tabacum* (Sugiyama et al., 2005), *Vitis vinifera* (Goremykin et al., 2009) and *Cucumis sativus* (Alverson et al., 2011a). In terms of mitochondrial genome size, the pigeonpea mitochondrial genome (545,742 bp) is substantially larger than the mitochondrial genome of closest sequenced legume species *Vigna radiata* (Alverson et al., 2011b) (401,262 bp). The size of pigeonpea mitochondrial genome was found to be comparable in size with the mitochondrial genomes of cereal species e.g. *O. sativa* (490 kb), *Z. mays* (569 kb) and greater than the median angiosperm mitochondrial genome size (473 kb). Mitochondrial genome size cannot only vary between different species but can also show variations between different lines of the same species. For instance, the genome size of five sequenced maize mitochondrial genomes is known to vary from 535,825 bp to 739,719 bp (Allen et al., 2007). The plant mitochondrial genomes are rich in non-coding regions and are highly variable in their non-coding regions. In our analysis, 12.29% of the pigeonpea mitochondrial genome was covered by the coding regions, which is lower than but comparable to *V. radiata* (Alverson et al., 2011b) (16.88%), *B. napus* (Chen et al., 2011) (17.34%) and *C. lanatus* (Alverson et al., 2010) (18.8%). However, the coding regions of the pigeonpea mitochondrial genome were found to be greater than the *C. pepo* in which only 6.9% (Alverson et al., 2010) of the mitochondrial genome is covered by coding regions. *C. pepo* has a particularly large mitochondrial genome (982,833 bp), due to the insertions of chloroplast (>113 kb) and short repeated sequences (> 370 kb) in the mitochondrial sequences (Alverson et al., 2010). The available plant mitochondrial genome sequences therefore suggest that gene composition does not depend on total genome size. Gene number can be conserved despite changes in size derived from the insertion of repetitive sequences, or of sequences transferred from the chloroplast and nucleus into inter-genic regions. In
addition, a large number of rearrangements of genes and ORFs are common features in plant mitochondrial genomes.
3.3.5 Comparison of gene order with other plant species

We also compared gene maps of protein-coding and rRNA genes encoded by the pigeonpea mitochondrial genome (ICPA 2039) with those of the other 11 sequenced plant species. Unsurprisingly, the highest level of synteny for the mitochondrial genomes was observed between C. cajan and the related legume V. radiata. For instance, one four-gene cluster (cox3-nad4L-atp4-rps10ab), two three-gene clusters (nad6-nad1-ccmB, rpl5-rps14-cob), and three two-gene clusters (rps3ab-rpl16, rps12-nad3, ccmC-ccmFn) were syntenic between these two species (Figure 3.2). On the other hand, only two two-gene clusters (nad3-rps12, ccmC-ccmFn) showed synteny in mitochondrial genomes of C. cajan with three cereals species analyzed (Supplementary Figure. 3.3, 3.4 and 3.5). Two two-gene clusters (nad2ab-atp1, rps3ab-rpl16) of the C. cajan mitochondrial genome showed synteny when compared with wheat and maize (Supplementary Figure 3.3 and 3.4). The most highly conserved gene cluster was the two-gene cluster of rps3ab-rpl16, which was syntenic across seven of the 11 species (Figure 3.2, Supplementary Figure 3.3–3.12).
Figure 3.2: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Vigna radiata*. Left hand side is represented by genes identified in *Cajanus cajan* and top side is represented by genes of *Vigna radiata*. Shaded blocks in the image represent the correlation of gene orders.
3.3.6 Comparison of mitochondrial genome sequences among Cajanus lines

Comparisons of gene order typically highlight the high rate of mitochondrial genome rearrangements between different plant species. Our analysis indicates that the mitochondrial genome of pigeonpea shares only six gene clusters with the closely related sequenced species V. radiata (Alverson et al., 2011b), which between them cover only 16 genes. Fewer gene clusters was observed to be conserved in comparison with more distantly related species. Due to the dynamic nature of the plant mitochondrial genomes, extensive structural variations can also be expected to occur in different lines of a given species (Davila et al., 2011). Comparison of mitochondrial genomes from five different lines of maize revealed 16 rearrangements, even between two fertile cytotypes (Allen et al., 2007). To understand the patterns of genetic variation associated with CMS in pigeonpea, and its effects on maternal inheritance, we first aligned the scaffolds of the mitochondrial genomes of the three Cajanus lines (i.e. ICPB 2039, ICPH 2433 and ICPW 29) using BLASTN, together, with that of the male-sterile line ICPA 2039. This demonstrated that the genomes of ICPA 2039 and ICPB 2039 are highly diverged from each other. Conversely, the mitochondrial genome of the hybrid ICPH 2433 (produced from the ICPA 2039 × ICPW 2433 cross) showed the highest level of synteny with ICPA 2039 line, followed by ICPW 29 (Figure 3.3), supporting the model that the CMS trait is maternally inherited in pigeonpea hybrids.

The sequence-level divergence between the mitochondrial genomes of ICPA 2039 and ICPB 2039 lines identified by BLASTN was further validated by use of the GS Reference Mapper 2.5, which is commonly used for mapping 454 reads to a reference assembly. Mapping raw reads of ICPB 2039, ICPH 2433 and ICPW 29 on to the assembly of ICPA 2039, a maximum number of rearrangements were observed in the fertile and sterile lines of ICPA 2039A system (Supplementary Figure 3.1). While building the comparative assemblies, we identified the no-coverage regions along with the rearrangements in order to reduce the effects of sequencing artifacts in further comparison. Twenty-two rearrangements and 17 no-coverage regions were observed in ICPB 2039 compared to ICPA 2039. Using the same criteria, nine rearrangements and 12 no-coverage regions were observed in the mitochondrial genome of ICPW 29 when compared to that of ICPA 2039. These nine
rearrangements could be the result of differences occurred during the maintenance of the CMS cytoplasm. We do not expect these differences to be associated with the CMS trait as the wild relative line is the maternal parent of the sterile line. The mitochondrial genome of ICPH 2433 was found to be closest in sequence to that of ICPA 2039, with no differences observed between these lines (Supplementary Figure 3.13, Supplementary Table 3.6 – 3.7).

3.3.7 Candidate CMS-associated chimeric ORFs

The CMS trait is often associated with chimeric open reading frames (ORFs) that are the products of mitochondrial genome rearrangement (Hanson and Bentolila, 2004) and which can cause pollen abortion. In some crops, chimeric ORFs are found in sterile lines (e.g. A-line) but absent in fertile lines (e.g. B- and R-lines) (Zabala et al., 1997; Hedgcoth et al., 2002; Allen et al., 2007). A number of studies have confirmed the role of chimeric ORFs in male sterility by disrupting the function of ORFs by inserting or deleting a few base pairs (reviewed by Hanson and Bentolila (Hanson and Bentolila, 2004)). Many of the chimeric genes associated with CMS in other crops are found in the proximity of protein-coding genes, and include regions encoding transmembrane domains and other parts of known mitochondrial genes. Hence, we set out to identify those chimeric ORFs which most closely resemble these criteria by scanning around positions that have undergone rearrangements or which are absent from particular pigeonpea lines. Only ORFs longer than 300 bp were considered, and were ranked as chimeric based on the presence of parts of other genes, proximity to known mitochondrial genes, and presence of hydrophobic domains. A scoring systems ranging from 0–4 was assigned to each ORF (see Chimeric ORFs section of Materials and methods). As abnormal atp synthase genes are sometimes associated with CMS (Young and Hanson, 1987; Kadowaki et al., 1990; Iwabuchi et al., 1993; Sabar et al., 2003), ORFs containing parts of atp genes were more heavily weighted.

Our study identifies 13 such potential CMS candidates in the pigeonpea male sterile line ICPA 2039 (Table 3.3). Of these 13 potential candidates, five carry parts of other mitochondrial genes and eight were observed to be in the proximity to other mitochondrial genes. Liu et al. (Liu et al., 2011) have hypothesized that a wheat K-type CMS line, Ks3, contains a chimeric ORF encoding partial subunits of several
components of the respiratory chain complex, including, *atp4*, *atp6*, *nad3*, *nad6*, *nad9*, *cox1* and *cox3*. These altered proteins may interfere with the normal function of respiratory chain reactions and cause pollen development to abort. Intriguingly, five of the candidates identified in our study incorporate parts of some of these genes including, *atp1*, *nad4*, *rps4*, *nad5* and *atp9*. This presents the possibility of a similar mechanism in pigeonpea CMS as found in rice. Transmembrane domains are another prominent feature associated with CMS ORFs (Bonnard and Grienenberger, 1995). Many of our candidate ORFs carry regions predicted to encode transmembrane domains and a number of the encoded proteins have been shown to be associated with the inner mitochondrial membrane (Hanson and Bentolila, 2004). Recently, Bentolila and Stefanov (Bentolila and Stefanov, 2012) have identified a candidate for the wild abortive-CMS in rice that has arisen via rearrangement, is chimeric in structure, possesses predicted transmembrane domains, as well as possess the promoter of a mitochondrial gene. Of our 13 candidates, 11 are predicted to carry such transmembrane domains. These novel ORFs may trigger CMS by damaging mitochondrial membrane structure such that the resulting permeability change affects mitochondrial function (Rhoads et al., 1995; Hanson and Bentolila, 2004). Previous histological studies of CMS in pigeonpea have revealed that meiosis in both male-fertile and male-sterile plants proceeds normally up to the tetrad stage, and that during this period, the tapetum remains intact. Male-sterility becomes manifest after this, with tetrads in male-sterile plants remaining enclosed within a persistent tetrad wall and subsequently undergoing vacuolation and abortion of pollen grains (Saxena et al., 2010a). Therefore, identifying the ORFs which are causative for CMS in pigeonpea will require the transcription and translation patterns of our unique ORF candidates to be determined, including in young to mature buds, floral parts including the pollen mother cell, tetrad and pollen grains. The roles of transmembrane domains and respiration in the mitochondrial genome of ICPA 2039 will also need to be assessed. Future, structural and functional studies will allow the exact mitochondrial genomic segments responsible for male-sterility in pigeonpea to be defined.
Figure 3.3: Alignments of *Cajanus* mitochondrial genomes. The outer circle represents the finalized mitochondrial genome assembly and gene annotation of male-sterile line ICPA 2039. Second, third, and fourth circle from the outer circle represent the scaffolds of ICPH 2433, ICPW 29, and ICPB 2039 mapped on ICPA 2039 assembly.
Accession numbers

Genome sequences and annotations from this article have been submitted to the GenBank data library under accession number SRA053693.
Table 3.3: Potential chimeric ORFs identified from the no coverage and rearrangement regions between ICPA 2039 and ICPB 2039 line.

<table>
<thead>
<tr>
<th>ORF start</th>
<th>ORF stop</th>
<th>ORF length</th>
<th>Nearest gene</th>
<th>Subject start</th>
<th>Subject stop</th>
<th>Chimera length</th>
<th>Identity</th>
<th>Subject features</th>
<th>No of transmembrane helices</th>
</tr>
</thead>
<tbody>
<tr>
<td>260331</td>
<td>260702</td>
<td>371</td>
<td>cox3</td>
<td>260342</td>
<td>260702</td>
<td>361</td>
<td>98</td>
<td>ORF 1</td>
<td>1</td>
</tr>
<tr>
<td>164867</td>
<td>165424</td>
<td>557</td>
<td>nad7</td>
<td>165424</td>
<td>165353</td>
<td>72</td>
<td>100</td>
<td>ORF 1</td>
<td>1</td>
</tr>
<tr>
<td>420342</td>
<td>420902</td>
<td>560</td>
<td>-</td>
<td>233322</td>
<td>232862</td>
<td>461</td>
<td>96</td>
<td>atp1 0</td>
<td>0</td>
</tr>
<tr>
<td>534744</td>
<td>535115</td>
<td>371</td>
<td>-</td>
<td>352424</td>
<td>352403</td>
<td>361</td>
<td>98</td>
<td>nad4 1</td>
<td></td>
</tr>
<tr>
<td>264435</td>
<td>265745</td>
<td>1310</td>
<td>-</td>
<td>60093</td>
<td>59813</td>
<td>281</td>
<td>100</td>
<td>rps4 2</td>
<td></td>
</tr>
<tr>
<td>165464</td>
<td>165853</td>
<td>389</td>
<td>nad7</td>
<td>265742</td>
<td>265981</td>
<td>241</td>
<td>97</td>
<td>ORF 3</td>
<td></td>
</tr>
<tr>
<td>164867</td>
<td>165424</td>
<td>557</td>
<td>nad7</td>
<td>165424</td>
<td>165353</td>
<td>72</td>
<td>100</td>
<td>ORF 1</td>
<td></td>
</tr>
<tr>
<td>276037</td>
<td>276405</td>
<td>368</td>
<td>-</td>
<td>468809</td>
<td>468752</td>
<td>58</td>
<td>98</td>
<td>atp9 3</td>
<td></td>
</tr>
<tr>
<td>396025</td>
<td>396876</td>
<td>851</td>
<td>mttB</td>
<td>264909</td>
<td>265396</td>
<td>488</td>
<td>97</td>
<td>ORF 2</td>
<td></td>
</tr>
<tr>
<td>396285</td>
<td>396641</td>
<td>356</td>
<td>mttB</td>
<td>265144</td>
<td>265396</td>
<td>253</td>
<td>99</td>
<td>ORF 1</td>
<td></td>
</tr>
<tr>
<td>264069</td>
<td>264434</td>
<td>365</td>
<td>-</td>
<td>44935</td>
<td>45092</td>
<td>158</td>
<td>100</td>
<td>nad5 1</td>
<td></td>
</tr>
<tr>
<td>165633</td>
<td>166088</td>
<td>455</td>
<td>nad7</td>
<td>265981</td>
<td>265886</td>
<td>96</td>
<td>100</td>
<td>ORF 0</td>
<td></td>
</tr>
<tr>
<td>8534</td>
<td>8842</td>
<td>308</td>
<td>ccmF</td>
<td>476033</td>
<td>476052</td>
<td>20</td>
<td>100</td>
<td>- 1</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4. Construction of a plant orthology database
4.1 Introduction

Current sequencing platforms are producing data with higher throughput and lower cost. As a result of this explosive growth of sequencing data, the number of uncharacterized sequences in public databases is increasing rapidly. Although experimental methods are the most accurate way of identifying protein function, they are labour intensive and time consuming. As a result, most genome scale functional annotations are based upon in silico predictions. Therefore, it is necessary to develop reliable methods and resources for the prediction of protein function. Homology prediction is one of the most successful approaches for protein function prediction.

The concept of orthology and paralogy was introduced to distinguish between two distinct types of homologous relationships. Orthologs are genes that are derived from a common ancestor through speciation, whereas paralogs are derived through a duplication event (Gabaldon and Koonin, 2013). Homology based function prediction is dependent on the ‘orthology conjecture’ hypothesis (Koonin, 2005; Studer and Robinson-Rechavi, 2009), although not without any controversy (Nehrt et al., 2011). As per the ‘orthology conjecture’ hypothesis, orthologs are expected to carry out identical (or more precisely, biologically equivalent) functions in different organisms, whereas paralogs tend to diverge after duplication to perform different functions via subfunctionalization and neofunctionalization (Gabaldon and Koonin, 2013).

The wide use of orthology prediction in genome annotation has promoted the development of several automated methods. The concept of orthology was originally defined in terms of pairwise comparisons between sequences, but in practice orthology clusters containing multiple species are more applicable in comparative genomics studies (Kristensen et al., 2011). Although a number of tools have been developed for the identification of orthologs, the best method is still under continuous debate. Large-scale ortholog identification is generally based on BLAST-based methods, as these methods have the advantage of being fast and scalable. I have used a BLAST-based method, orthoMCL (Barrier et al., 2003) for the identification of orthology clusters in 34 sequenced Viridiplantae species. OrthoMCL has previously been shown to have the best performance (and the lowest false positive rate) in ortholog detection across multiple genomes (Chen et al., 2007; Datta et al., 2009). The OrthoMCL algorithm uses reciprocal best BLAST hits (i.e., the first sequence
finds the second sequence as its best hit in the second species, and vice versa) and clusters closely related proteins based on Markov CLuster algorithm (MCL). The OrthoMCL algorithm clusters ‘recent paralogs’ (those which duplicated after speciation) together with the orthologs, as it is assumed that genes that have evolved from ancient duplication events (duplication after speciation) may have diverged to evolved new functions whereas, recent duplicated genes are likely to retain same functions and should be grouped with true orthologs (Barrier et al., 2003). Therefore, an orthoMCL cluster represents orthologous genes identified between different species plus recently evolved paralogs.

I developed a plant orthology database due to the underrepresentation of plant genomes in existing orthology databases (at the time of construction of this database in 2011). For example, GreenPhylDBv2.0, released in 2011, included 16 plant species, whereas more than 30 fully annotated plant genomes were available. Our orthology database includes 34 Viridiplantae species, 32 species of land plant (Embryophyte) and two green algae species (Chlorophyte). The 32 land plant species includes major plant clades, such as Brassicaceae, Poaceae (or grasses) and Fabaceae (Table 4.1). This database has been developed as a community resource and for identification of signatures of positive selection on A. thaliana imprinted genes (Chapter 5).
4.2 Materials and Methods

4.2.1 Data

Peptide sequences of all the sequenced species available in Phytozome database were downloaded (Phytozome 8.0; (Goodstein et al., 2012)). *Cajanus cajan* peptide sequences were downloaded from the supplementary material provided with the publication (Varshney et al., 2011). *Lotus japonicus* peptide sequences were downloaded from the LjGDB database (http://www.plantgdb.org/LjGDB/). Table 4.1 list the 34 species used for constructing the orthology database, their annotation version and their clades.

4.2.2 Construction of orthology database

OrthoMCL (Li et al., 2003) standalone version was downloaded from OrthMCL database website (http://orthomcl.org/common/downloads/software/version 2.0). The longest transcript was used as the representative transcript for each gene. Transcripts for each gene were quality checked to ensure that the CDS has no internal STOP codons. The Irish Centre for High-End Computing (ICHEC) facility was used to perform all-vs-all BLASTP on peptide sequences of 34 species. In order to minimize the number of false positives and ensure tight clustering of genes families, we detected these orthologous relationships between sequences using OrthoMCL (Li et al., 2003). Figure 4.1 represents the pipeline constructed for identification of orthology clusters. OrthoMCL algorithm has three major steps: 1) performing an all-against-all BLAST (Altschul et al., 1990); 2) identification of putative orthology and inparalogy relations. An expectation value of $e^{-5}$ and a percent match of 50 were applied as cut off criteria for orthology and inparalogy relationship identification; 3) generation of disjoint clusters of closely related proteins with the graph clustering algorithm MCL (Enright et al., 2002). The inflation value parameter of the MCL algorithm was used to control the tightness of putative clusters. An inflation value of 1.5 was used, as this has been defined as the optimal cutoff by Li et al. (2003).
Table 4.1: List of 34 sequenced plant and algal species that were used to construct plant orthology database.

<table>
<thead>
<tr>
<th>Species Name</th>
<th>Code</th>
<th>Common name</th>
<th>Clades’</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>Ath</td>
<td>Thale cress</td>
<td>Br, Eu, Ag, Em, Vi</td>
<td>TAIR10</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>Aly</td>
<td>Lyrate rockcress</td>
<td>Br, Eu, Ag, Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>Bra</td>
<td>Turnip mustard</td>
<td>Br, Eu, Ag, Em, Vi</td>
<td>v1.2</td>
</tr>
<tr>
<td>Capsella rubella</td>
<td>Cru</td>
<td>Red shepherd</td>
<td>Br, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td>Eutrema salsugineum</td>
<td>Tha</td>
<td>Salt cress</td>
<td>Br, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td>Aquilegia coerulea</td>
<td>Aco</td>
<td>Columbine</td>
<td>Eu, Ag, Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td>Carica papaya</td>
<td>Cpa</td>
<td>Papaya</td>
<td>Eu, Ag, Em, Vi</td>
<td>v0.4</td>
</tr>
<tr>
<td>Citrus Sinensis</td>
<td>Csi</td>
<td>Sweet orange</td>
<td>Eu, Ag, Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td>Citrus clementina</td>
<td>Ccl</td>
<td>Clementine</td>
<td>Eu, Ag, Em, Vi</td>
<td>v0.9</td>
</tr>
<tr>
<td>Eucalyptus grandis</td>
<td>Egr</td>
<td>Eucalyptus</td>
<td>Eu, Ag, Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>Vvi</td>
<td>Grape vine</td>
<td>Eu, Ag, Em, Vi</td>
<td>Genoscope.12X</td>
</tr>
<tr>
<td>Mimulus guttatus</td>
<td>Mgu</td>
<td>Monkey flower</td>
<td>Eu, Ag, Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td>Manihot esculenta</td>
<td>Mes</td>
<td>Cassava</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v4.1</td>
</tr>
<tr>
<td>Ricinus communis</td>
<td>Rco</td>
<td>Castor bean</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v0.1</td>
</tr>
<tr>
<td>Linum usitatissimum</td>
<td>Lus</td>
<td>Flax</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>Ptri</td>
<td>Western poplar</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v2.2</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Mtr</td>
<td>Barrel medic</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v3</td>
</tr>
<tr>
<td>Phaseolus vulgaris</td>
<td>Pvu</td>
<td>Common bean</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v0.91</td>
</tr>
<tr>
<td>Glycine max</td>
<td>Gma</td>
<td>Soybean</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td>Scientific Name</td>
<td>Abbreviation</td>
<td>Common Name</td>
<td>Major Clades</td>
<td>Version</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------</td>
<td>-------------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td><em>Cucumis sativus</em></td>
<td>Csa</td>
<td>Cucumber</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td><em>Prunus persica</em></td>
<td>Ppe</td>
<td>Peach</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td><em>Malus domestica</em></td>
<td>Mdo</td>
<td>Apple</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td><em>Lotus japonicus</em></td>
<td>Lja</td>
<td>Miyakogusa</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>Lj1.0</td>
</tr>
<tr>
<td><em>Cajanus cajan</em></td>
<td>Cca</td>
<td>Pigeonpea</td>
<td>Fa, Eu, Ag, Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td><em>Brachypodium distachyon</em></td>
<td>Bdi</td>
<td>Purple brome</td>
<td>Po, Ag, Em, Vi</td>
<td>v1.2</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em></td>
<td>Sbi</td>
<td>Cereal grass</td>
<td>Po, Ag, Em, Vi</td>
<td>v1.4</td>
</tr>
<tr>
<td><em>Panicum virgatum</em></td>
<td>Pvi</td>
<td>Switchgrass</td>
<td>Po, Ag, Em, Vi</td>
<td>v0.0</td>
</tr>
<tr>
<td><em>Zea mays</em></td>
<td>Zma</td>
<td>Maize</td>
<td>Po, Ag, Em, Vi</td>
<td>v6a</td>
</tr>
<tr>
<td><em>Setaria italic</em></td>
<td>Sit</td>
<td>Foxtail millet</td>
<td>Po, Ag, Em, Vi</td>
<td>v2.1</td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>Osa</td>
<td>Rice</td>
<td>Po, Ag, Em, Vi</td>
<td>v6.0</td>
</tr>
<tr>
<td><em>Selaginella moellendorffii</em></td>
<td>Smo</td>
<td>Spikemoss</td>
<td>Em, Vi</td>
<td>v1.0</td>
</tr>
<tr>
<td><em>Physcomitrella patens</em></td>
<td>Ppa</td>
<td>Moss</td>
<td>Em, Vi</td>
<td>v1.1</td>
</tr>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>Cre</td>
<td>Green Algae</td>
<td>Ch, Vi</td>
<td>v4.3</td>
</tr>
<tr>
<td><em>Volvox carteri</em></td>
<td>Vca</td>
<td>Green Algae</td>
<td>Ch, Vi</td>
<td>v2.0</td>
</tr>
</tbody>
</table>

*Major plant and algal clades: Br- Brassicaceae, Fa- Fabidae, Eu- Eudicot, Po- Poaceae, Ag- Angiosperms, Em- Embryophyte, Ch- Chlorophyte, Vi- Virdiplantae.*
Figure 4.1: Pipeline for identification of orthology clusters in sequenced plant species. Orange boxes show data information, grey boxes show the programs from the OrthoMCL pipeline and green boxes display the programs developed in-house.
4.3 Results and example uses of the database

4.3.1 Description of the orthology database

Orthomcl was used to cluster 1,162,324 protein sequences from 34 sequenced Viridiplantae species. These sequences were clustered into 76,234 orthology clusters representing 905,992 (77.9%) protein sequences, leaving 256,332 (22.05%) singleton genes (Table 4.2).

Table 4.2: Statistics of orthology database of 34 sequenced species.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of protein sequences</td>
<td>1,162,324</td>
</tr>
<tr>
<td>Number of orthology clusters</td>
<td>76,234</td>
</tr>
<tr>
<td>Number of sequences clustered in orthology</td>
<td>905,992</td>
</tr>
<tr>
<td>Number of singletons</td>
<td>256,332</td>
</tr>
</tbody>
</table>

4.3.2 Using the orthology database

The orthology database is kept in a simple text file format, as this is easy to use compared to a mysql database. In a Linux based system database can be used simply with a grep command. Below I am representing an example of the extraction of orthologs for the A. thaliana gene AT2G47520. Phyto represents the cluster id assigned to an orthology cluster. The initial three letters (separated from geneid by ‘|’) represent the three-letter species code assigned to each species. The orthology cluster for AT2G47520 represents three species and 11 sequences.

Command: grep 'AT2G47520' Orthologydatabase_34species

4.4 Conclusions

Due to the underrepresentation of plant species in existing orthology database, I have developed an orthology database in this chapter. This database contains orthology clusters identified in peptide sequences of 32 land plant species and 2 algae species. It has been used as an internal lab resource for identification of orthologs between different species for various lab projects. I have used this database in the next chapter for identification of signatures of positive selection in *A. thaliana* imprinted genes.
Chapter 5. Signatures of positive Darwinian selection in *Arabidopsis thaliana* paternally expressed imprinted genes
5.1 Introduction

Land plant life cycles differ from animals due to the presence of multicellular diploid and haploid generations termed the sporophyte and gametophyte respectively, in both of which gene expression occurs. It has been predicted that this may alter the forms of selective pressure to which plant genomes are subject (Walbot and Evans, 2003). Tests of this hypothesis have demonstrated that genes expressed in the highly reduced male gametophyte of flowering plants (the pollen grain), display elevated levels of positive selection. This phenomenon has been described in the highly outcrossing species Capsella grandiflora (Arunkumar et al., 2013) but also in Arabidopsis thaliana, which is predominantly inbreeding (Gossmann et al., 2014). In both cases, increased positive selection is observed in genes expressed in the pollen tube but not in those expressed in the sperm cell. This has been interpreted as a consequence of sexual selection due to competition between the pollen grains for access to ovules.

Functionally haploid gene expression also occurs in the flowering plant endosperm at loci, which are subject to genomic imprinting (Garnier et al., 2008; Kohler et al., 2012). Imprinting is an epigenetic phenomenon which violates the usual rules of epigenetic inheritance, leading to a gene being monoallelically expressed in a parent of origin-dependent manner. Interestingly, in genus Arabidopsis positive selection has been shown to occur at the MEDEA locus, which is only expressed from the maternal genome in the endosperm due to imprinting (Spillane et al., 2007). Positive selection on imprinted loci has been interpreted as a signature of conflict between the maternal and paternal genomes over allocation of resources from the sporophyte of the maternal plant (the parental conflict theory, after Haig and Westoby (1991)). Positive selection at MEDEA has so far only been demonstrated in the species Arabidopsis lyrata, which is predominantly outcrossing while this selective pressure has been lost in A. thaliana. However, a study of one group of imprinted genes in A. thaliana suggested that they display accelerated evolutionary rates compared to the genomic background as indicated by elevated $dN/dS$ ratios, which measure the relative frequency of synonymous and non-synonymous nucleotide substitutions (Wolff et al., 2011a).
Evolutionary fate of newly evolved mutations is determined by random genetic drift, positive selection or purifying selection. Synonymous mutations do not change the amino acid whereas nonsynonymous mutations lead to change in the amino acid. Therefore the substitution rate of the nonsynonymous mutations \((dN)\) is function of selection pressure on the protein product of a gene while substitution rate of the synonymous mutations is neutral. The ratio of nonsynonymous substitutions per nonsynonymous site \((dN)\) to synonymous substitutions per synonymous site \((dS)\) is known as \(\omega\), and is an indication of selection pressure. If \(\omega > 1\) signifies positive selective pressure, \(\omega = 1\) signifies neutral evolution, while \(\omega < 1\) indicates purifying selective pressure. I have conducted this study to test the evidence of positive Darwinian selection in *Arabidopsis thaliana* imprinted genes (iPEGs and iMEGs) by applying Maximum likelihood methods that are based on codon models of sequence evolution. These methods were considered to be robust than other methods of detecting positive selection like sliding window approach (Schmid and Yang, 2008).

This analysis is conducted on orthologs of imprinted genes by utilizing the plant orthology database (described in Chapter 4) that contains the orthology clusters of 34 sequenced Viridiplantae species. As the power of Maximum likelihood methods increase with greater taxonomy representation and greater divergence (Anisimova et al., 2001), I have used the imprinted genes retaining > 5 species in orthology clusters to achieve statistical confidence in our results. In this chapter, I have explored the following questions regarding the evolution of imprinted genes: 1) are any *A. thaliana* imprinted genes evolving under positive Darwinian selection? 2) does iPEGs experience different level of selective pressure compared to iMEGs? 3) does imprinted genes (iPEGs or iMEGs) experience different levels of positive selection compared to the genes randomly selected from *A. thaliana* genome? 4) does imprinted genes (iPEGs or iMEGs) experience different levels of positive selection compared to the genes randomly selected from set of *A. thaliana* genes specific to endosperm (Belmonte et al., 2013)?

Further I have used *A. thaliana* population data (Cao et al., 2011) to detect signatures of selection pressure in imprinted genes with \(\leq 5\) species (app. 50% of our total imprinted genes set) in their orthology clusters. Positive selection may also drive the
fixation of beneficial traits within a population (Sabeti et al., 2006). Therefore, *A. thaliana* population data is also used to identify the level of fixation of positively selected sites identified using codeml.
5.2 Materials and methods

5.2.1 Data

A total of 436 *A. thaliana* imprinted gene set was compiled from a number of recent high-throughput screens (Gehring et al., 2011; Hsieh et al., 2011; McKeown et al., 2011; Wolff et al., 2011b) and other studies (Vielle-Calzada et al., 1999; Kinoshita et al., 2004; Kohler et al., 2005; Jullien et al., 2006; Tiwari et al., 2008; Gehring et al., 2009). Genes that were not experimentally validated were discarded from further analysis.

5.2.2 Gene age and evolutionary rate

BLASTP of imprinted genes against all the sequenced plant species was conducted to predict the gene age (e-value <1e-3). Age class is assigned to each imprinted gene based on BLASTP hits. AC 0 represents the youngest age class and AC 9 as the oldest class. Age Class (AC): AC0- *A. thaliana* specific, AC1- *A. lyrata*, AC2- Brassicaceae, AC3- Brassicales-Malvales, AC4- Rosid, AC5- Eudicot, AC6- Angiosperm, AC7- Tracheophyte, AC8- Embryophyte, AC9- Viridiplantae. Pair wise $dN/dS$ calculations between reciprocal best hits of *A. thaliana* and *A. lyrata* were adapted from Donoghue and Spillane (2010). Statistical test was preformed as implemented in R package (r-project.org).

5.2.3 Orthologs Identification

Orthologs of 148 imprinted genes were extracted from the plant orthology database (Chapter 4). As the power of Maximum likelihood methods increase with greater taxonomy representation and greater divergence (Anisimova et al., 2001), imprinted genes with orthologs in > 5 species and orthology cluster size > 6 were used for codeml analysis to achieve statistical significance in results.
5.2.4 Random Sets

A random set of 100 genes was generated from *A. thaliana* genome. Along with the random set from genome, a 100 genes random set was also generated from endosperm-specific genes (Belmonte et al., 2013). Both random sets were generated from genes where the genes retained > 5 species and orthology cluster size > 6 in their orthology cluster.

5.2.5 Multiple Sequence Alignment

Multiple sequence alignments for each orthology cluster were constructed using AQUA (Muller et al., 2010). AQUA uses MUSCLE (Edgar, 2004) and MAFFT (Katoh and Toh, 2008) as alignment programs, RASCAL (Thompson et al., 2003) for refining the alignment and norMD score (Thompson et al., 2001) for assessment of their quality. Alignment qualities were assessed using norMD score (Thompson et al., 2001). Alignments with a norMD score <0.6 are referred to as low quality (Thompson et al., 2001). Any specific sequences within these alignments that were poorly aligned were removed and the norMD score was recalculated. If the norMD score subsequently increased above the threshold of 0.6 the alignment was retained for further analysis. The nucleotide sequence alignments were generated for each family using the amino acid alignment and original nucleotide sequence files, this was achieved using a python script (from Mary O’Connell lab).

5.2.6 Recombination Detection

The recombinant sequences were identified using RDP3 (Martin et al., 2010). We used two substitution-based methods GENECONV (Sawyer, 1989) and MaxChi (Smith, 1992) and two phylogenetic-based methods BOOTSCAN (Martin et al., 2005) and SiScan (Gibbs et al., 2000) for identification of recombinant sequences. A sequence is considered to be a recombinant and removed from further analysis if the recombination event was significant for a substitution-based method as well as for a phylogenetic-based method.
5.2.7 Tree building

Models for protein sequence evolution were generated using the program modelgenerator (Keane et al., 2006). Phylogenetic trees were inferred using RAxML version 7.2.6 (Randomized Axelerated Maximum Likelihood) (Stamatakis, 2006) with 1000 bootstrap replicates. Rapid bootstrapping algorithm was used in RAxML to create the phylogenetic trees. Due to the memory restrictions on ICHEC server the codeml analysis was conducted on clade of interest for genes with greater than 80 sequences in their orthology clusters.

5.2.8 Selection Analysis

Selection analysis was conducted using the PAML version 4.4e (Yang, 2007). Both lineage-specific models and site-specific models were evaluated using likelihood ratio test (LRT) (Yang, 1998; Yang et al., 2000; Yang and Nielsen, 2002). Yang and Nelson (2002) codon substitution models were used to identify selection pressure acting on different sites along specified lineages. A sequence is considered to be under lineage-specific analysis if the likelihood ratio test was significant for comparison of both ModelA null and M1Neutral models with ModelA. M1Neutral is a neutral model that allows two site classes for proportion of sites where $\omega_0 = 0$ and $\omega_1 = 1$. Model A assumes two site classes are same in both foreground and background lineages ($\omega_0=0$ and $\omega_1=1$) and $\omega_2$ is calculated. Model A null is the null hypothesis for this model that allows for sites evolving under purifying selection, or neutrally evolving in the background lineages.

For site-specific analyses, likelihood ratio tests were conducted to compare models M7 and M8a with model M8. The test compares the neutral model M7, which assumes a $\beta$ distribution for $\omega$ over sites and the alternative model M8 ($\beta$ and $\omega$), which adds an extra site class of positive selection. M8a is the null hypothesis of M8 where the additional category is neutral, i.e. $\omega=1$. CodeML wrapper (software from Mary O’Connell lab) was used to prepare all the codeML files, to parse the PAML output and perform the likelihood ratio test.
After ML estimates of model parameters are obtained, we used the empirical Bayes approach to infer the posterior probability of the positively selected sites using two approaches: Bayes Empirical Bayes (BEB) and Naïve Empirical Bayes (NEB). BEB approach reduces the rate of false positives when analyzing small datasets and retains the good power of NEB approach when analyzing large datasets (Yang et al., 2005). Therefore results from BEB calculations were preferred, if both NEB and BEB are predicted. Further, we discarded any predicted positive selected sites overlapping with the poor alignment regions (gaps > 40%). TrimAL (Capella-Gutierrez et al., 2009) was used to calculate the percentage of gaps in the alignments (-sgc option).

5.2.9 Population Genetic Analysis

*A. lyrata* ortholog pairs of *A. thaliana* imprinted genes were identified using reciprocal best hits (RBH). A total of 118 imprinted genes showed the reciprocal best hits with the *A. lyrata* genes. *A. thaliana* and *A. lyrata* CDS alignments were created using the method described for multiple sequence alignments in section 5.2.4. *A. thaliana* 80 accessions SNP data (Cao et al., 2011) was downloaded for the 1001 genome project (http://1001genomes.org/data/MPI/MPICao2010/releases/current/genome_matrix/). Multiple accessions SNP data was mapped to the *A.thaliana* using a custom-made python script.

McDonald-Kreitman test on each of the imprinted gene was performed using a python script that uses egglib library to calculate $D_n$, $D_s$, $P_n$ and $P_s$ values. Fisher's exact test was used to test the statistical significance of results. Fixation index ($FI$) were determined as $FI= (D_n/D_s)/(P_n/P_s)$. Expected fixation index ($eFI$) is calculated as given in Axelsson and Ellegren, 2009.
5.3 Results

5.3.1 iPEGs and iMEGs showed similar age distribution and evolutionary rates, but accelerated evolutionary rates compared to whole genome

A set of 436 *Arabidopsis thaliana* imprinted genes (435 literature+1 unpublished) were compiled from all relevant high-throughput screens (Gehring et al., 2011; Hsieh et al., 2011; McKeown et al., 2011; Shirzadi et al., 2011; Wolff et al., 2011b) and other studies (Vielle-Calzada et al., 1999; Kinoshita et al., 2004; Kohler et al., 2005; Jullien et al., 2006; Tiwari et al., 2008; Gehring et al., 2009). These genes have been predicted to be uniparentally-expressed due to genomic imprinting in *A. thaliana* seeds. Genes that had not been experimentally validated as uniparentally-expressed were discarded from further analysis and a validated set of 148 uniparentally-expressed genes was considered for selection analysis in this study (Supplementary Table 5.1).

Firstly, the age of 148 imprinted genes were determined using a phylostratigraphy approach. Nine phylogenetic classes (Age class (AC)) were defined for the sequenced plant species in Phytozome database (http://www.phytozome.net/) where AC 0 represents the youngest age class and AC 9 as the oldest class. Using an e-value cutoff of $< 10^{-3}$, the imprinted genes were classified in different age classes (Figure 5.1). No significant difference was observed between the age distributions of iMEGs and iPEGs. A total of 7 imprinted genes did not show any sequence similarity outside Brassicaceae i.e. Brassicaceae-specific orphans in *A. thaliana* (identified by Donoghue et al., 2011). Out of these 7 imprinted *Arabidopsis* specific orphans, 1 gene was identified as *A. thaliana* specific orphan.

Using *A. lyrata* orthologs of these imprinted genes, the rate of protein evolution of imprinted genes was identified. Both iPEGs and iMEGs showed omega significantly higher than that of the background (Table 5.1; U-test, iPEGs P=9.9e-07, iMEGs P=1.9-06). No significant difference was observed between the omega value of iPEGs and iMEGs.
Figure 5.1: Distribution of age of iMEGs and iPEGs. Age Class (AC): AC0- A. thaliana specific, AC1- A. lyrata, AC2- Brassicaceae, AC3- Brassicales-Malvales, AC4- Rosid, AC5- Eudicot, AC6- Angiosperm, AC7- Tracheophyte, AC8- Embryophyte, AC9- Viridiplantae.

Table 5.1: Evolutionary rate of imprinted genes in comparison to the whole genome as background.

<table>
<thead>
<tr>
<th></th>
<th>Mean (\omega (dN/dS))</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPEGs</td>
<td>0.4361±0.053</td>
</tr>
<tr>
<td>iMEGs</td>
<td>0.5045±0.061</td>
</tr>
<tr>
<td>Background</td>
<td>0.2437±0.002</td>
</tr>
</tbody>
</table>
5.3.2 *A. thaliana* imprinted genes show enrichment for recent whole genome duplication events (alpha)

Orthology clusters of 148 imprinted genes were fetched from the plant orthology database created using OrthoMCL (Chapter 4). Imprinted genes with cluster size > 6 and > 5 species in their orthology clusters were considered for codeml analysis to achieve the statistical confidence in results.

A total of 72 imprinted genes (~50%) were found to be in less than six species. Figure 5.2 represents the number of taxa in imprinted gene families. Out of these 72 imprinted genes 32 genes are iMEGs and 40 iPEGs (Supplementary Table 5.2). Two genes (AT4G11400 an iPEG and AT5G53870 an iMEG) were found to be of poor sequence alignment quality (norMD score < 0.6) and were excluded from the codeML analysis. These two genes were tested for selective pressure using McDonald-Kreitman analysis. Except one gene (AT4G31060) all the imprinted genes were found to contain multiple genes in their orthology clusters.

The 148 imprinted gene set has shown enrichment for participation in the alpha whole genome duplication event (52 imprinted genes in alpha events, fisher test, P=0.02), whereas, only 21 genes were found to participated in beta and gamma duplication events (fisher test, P=0.14). The alpha duplication event is predicted to be the most recent Brassicaceae-specific WGD event (Franzke et al., 2011) while the beta and gamma events are predicted to be older (Bowers et al., 2003). The summarized data indicates that very few imprinted genes were originated from the older duplication events. These findings are in agreement with Qiu et al., which showed that a large number of imprinted genes were formed by duplication during the evolution of Brassicales (Qiu et al., 2014).
Figure 5.2: Number of species in orthology clusters of imprinted genes. *All species on X-axis represents all 34 Viridiplante species from phytozome database. Y-axis represents the number of gene families.*
5.3.3 Imprinted genes are under positive selection

For studying the evolution of imprinted genes, 148 imprinted genes were divided into two categories: a) imprinted genes with orthologs in > 5 species and > 6 sequences in their orthology clusters, b) imprinted genes in ≤ 5 species and ≤ 6 sequences in their orthology clusters (Figure 5.3). A total of 7 Brassicaceae-specific orphan imprinted genes in our dataset.

Figure 5.3: Schematic for identification of evolution of 148 A. thaliana imprinted genes. Level of positive selection in a) 70 imprinted genes were assessed using codeml pipeline, b) 64 imprinted genes were studied using McDonald-Kreitman test.
Using the codeml models (as described in section 5.2.7), I have identified a number of imprinted genes under lineage-specific (Table 5.3) and site-specific selection (Table 5.4). Interestingly, a total of 11 imprinted genes were identified to be both under lineage specific selection and site-specific selection (Figure 5.4). A higher percentage of number of positive selected genes is observed in iPEGs compared to iMEGs (Table 5.2). However, no statistical significance was observed for this difference both for lineage-specific selection and site-specific selection (fisher-test, P=0.1 and P=0.6).

Figure 5.4: Venn diagram for imprinted genes found be both under lineage-specific selection and site-specific selection using codeML models.
Table 5.2: Number of iMEGs and iPEGs identified to be under lineage-specific selection and site-specific using codeML models.

<table>
<thead>
<tr>
<th></th>
<th>iMEGs</th>
<th>iPEGs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of genes</td>
<td>31</td>
<td>39</td>
<td>70</td>
</tr>
<tr>
<td>Lineage-specific selection</td>
<td>6 (19.4%)</td>
<td>15 (38.5%)</td>
<td>21 (30.0%)</td>
</tr>
<tr>
<td>Site-specific selection</td>
<td>11 (35.5%)</td>
<td>17 (43.6%)</td>
<td>28 (40.0%)</td>
</tr>
</tbody>
</table>
Table 5.3: Sites in imprinted genes identified to be under lineage-specific positive selection.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Imprinting Status</th>
<th>Positively Selected Sites</th>
<th>P-value of positive selected sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT2G28380</td>
<td>iMEG</td>
<td>272*, 361, 367, 368, 370, 372, 374, 402, 407, 414</td>
<td>$10 &gt; 0.5, 1 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>iMEG</td>
<td>53, 108*, 136, 178*, 181, 187, 193, 225*, 235, 240, 247, 273, 276, 291, 475, 581*</td>
<td>$16 &gt; 0.5, 4 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>iMEG</td>
<td>170, 172, 173, 176, 186, 227, 234, 238, 239, 240*, 241, 250, 342*, 422, 423, 439, 441*, 455, 459, 929, 1017, 1047, 1101, 1145</td>
<td>$24 &gt; 0.5, 3 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>iMEG</td>
<td>42, 338, 622, 1024, 1187, 1210, 1221, 1240*</td>
<td>$8 &gt; 0.5, 1 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G57820</td>
<td>iPEG</td>
<td>218, 442, 565, 592, 605, 622</td>
<td>$6 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G48910</td>
<td>iPEG</td>
<td>153*, 362</td>
<td>$2 &gt; 0.5, 1 &gt; 0.95, 1 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G63020</td>
<td>iPEG</td>
<td>77, 201, 932, 1177, 1287</td>
<td>$5 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G67820</td>
<td>iPEG</td>
<td>239, 404*, 410*</td>
<td>$3 &gt; 0.5, 2 &gt; 0.95, 1 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G53150</td>
<td>iPEG</td>
<td>49, 62, 115, 150, 439, 442, 509, 526, 624</td>
<td>$9 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G49600</td>
<td>iPEG</td>
<td>105, 349*</td>
<td>$2 &gt; 0.5, 1 &gt; 0.95, 1 &gt; 0.99$</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>iPEG</td>
<td>5, 29, 53, 62, 63, 68, 77, 82, 100, 102, 103, 121, 124, 128, 132, 135, 145, 148, 179*, 184, 233, 234, 246, 285*, 287, 330, 333, 356, 357*, 381*, 386, 395, 423, 427, 431, 432, 442, 478, 479</td>
<td>$39 &gt; 0.5, 4 &gt; 0.95, 2 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G47490</td>
<td>iPEG</td>
<td>20, 115, 125, 173, 231, 321, 392</td>
<td>$7 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G28300</td>
<td>iPEG</td>
<td>72, 575, 576, 578, 579, 581</td>
<td>$6 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G59890</td>
<td>iPEG</td>
<td>295, 1025</td>
<td>$2 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99$</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Type</td>
<td>Features Reference</td>
<td>Score Details</td>
</tr>
<tr>
<td>-------------</td>
<td>------</td>
<td>--------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>AT1G20910</td>
<td>iPEG</td>
<td>23, 84, 93, 97*, 113, 114, 123, 139*, 147, 153, 169, 179*, 183, 195*, 198*, 234, 255, 257, 269, 270, 308*, 317, 322, 335, 339, 346*, 347, 368, 372, 384, 386, 392</td>
<td>32 &gt; 0.5, 7 &gt; 0.95, 2 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G61330</td>
<td>iPEG</td>
<td>15, 32, 300, 429</td>
<td>4 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G17770</td>
<td>iPEG</td>
<td>34, 104, 141, 157, 163, 211, 252, 254, 263, 330, 349, 354, 437, 570, 693</td>
<td>15 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G54280</td>
<td>iMEG</td>
<td>965*</td>
<td>1 &gt; 0.5, 1 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT2G32370</td>
<td>iPEG</td>
<td>703 sites (Supplementary Table 5.6)</td>
<td>703 &gt; 0.5, 3 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
</tbody>
</table>

* Positively selected sites with posterior probability > 0.95 are marked with asterisk (*)
Table 5.4: Sites in imprinted genes identified to be under site-specific positive selection.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Imprinting Status</th>
<th>Positively Selected Sites</th>
<th>P-value of positive selected sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G18650</td>
<td>iMEG</td>
<td>228, 232</td>
<td>2 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>iMEG</td>
<td>537, 1038, 1124, 1130, 1232</td>
<td>5 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT5G03370</td>
<td>iMEG</td>
<td>33, 36, 37, 48, 60, 64, 77</td>
<td>7 &gt; 0.5, 1 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G76250</td>
<td>iMEG</td>
<td>54, 74, 77, 78, 79, 80</td>
<td>6 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>iMEG</td>
<td>299, 315, 1245, 1254, 1260, 1261</td>
<td>6 &gt; 0.5, 1 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G62660</td>
<td>iMEG</td>
<td>14, 62, 63, 64, 65*, 66, 86*</td>
<td>7 &gt; 0.5, 2 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT3G14205</td>
<td>iPEG</td>
<td>5, 9, 10, 12, 14, 17, 450</td>
<td>7 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G48910</td>
<td>iPEG</td>
<td>379*</td>
<td>1 &gt; 0.5, 1 &gt; 0.95, 1 &gt; 0.99</td>
</tr>
<tr>
<td>AT3G45090</td>
<td>iPEG</td>
<td>20, 76, 78, 82, 83, 87, 88, 91, 623*, 624*, 629*, 636, 638, 685</td>
<td>14 &gt; 0.5, 3 &gt; 0.95, 1 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G67820</td>
<td>iPEG</td>
<td>30</td>
<td>1 &gt; 0.5, 1 &gt; 0.95, 1 &gt; 0.99</td>
</tr>
<tr>
<td>AT5G11460</td>
<td>iPEG</td>
<td>6, 182, 254, 265</td>
<td>4 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G70560</td>
<td>iPEG</td>
<td>3, 10, 11, 13, 16, 18*, 19*, 385, 386*</td>
<td>9 &gt; 0.5, 3 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>Gene Name</td>
<td>Type</td>
<td>Accession Numbers</td>
<td>P-values</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>AT3G08040</td>
<td>iPEG</td>
<td>10, 14, 109, 112, 114, 116, 134, 136, 137, 139, 140, 141, 143, 144, 145, 147, 148, 149, 150, 151, 152, 153, 154, 156, 158, 160, 213, 291, 520, 521</td>
<td>$3 &gt; 0.5$, $5 &gt; 0.95$, $1 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G53150</td>
<td>iPEG</td>
<td>431, 439, 695</td>
<td>$3 &gt; 0.5$, $0 &gt; 0.95$, $0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G37290</td>
<td>iPEG</td>
<td>180*</td>
<td>$1 &gt; 0.5$, $1 &gt; 0.95$, $0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G43780</td>
<td>iPEG</td>
<td>9, 11, 12, 13*, 14, 15, 17*, 19, 20*, 21*, 22*, 23, 24*, 25*, 26*, 27*, 28, 29*, 32, 33*, 34*, 36*, 37, 43*, 46*</td>
<td>$25 &gt; 0.5$, $15 &gt; 0.95$, $5 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G42670</td>
<td>iPEG</td>
<td>45, 61, 63, 96, 161, 167, 212, 214, 226, 242, 273</td>
<td>$11 &gt; 0.5$, $0 &gt; 0.95$, $0 &gt; 0.99$</td>
</tr>
<tr>
<td>AT5G28300</td>
<td>iPEG</td>
<td>228*, 233, 248, 382, 383, 384, 387, 406, 407, 408, 415, 424, 426*, 427, 428, 429*, 430, 440, 571, 576, 577*, 578*, 579*, 588, 592, 596</td>
<td>$26 &gt; 0.5$, $6 &gt; 0.95$, $3 &gt; 0.99$</td>
</tr>
<tr>
<td>AT1G20910</td>
<td>iPEG</td>
<td>7, 8, 11*, 12, 17, 19, 22, 23*, 24, 25, 27, 28*, 35, 39, 43, 48, 49, 50, 52, 53, 54, 57, 62, 65, 69, 71, 72, 73, 75, 76*, 398</td>
<td>$31 &gt; 0.5$, $4 &gt; 0.95$, $3 &gt; 0.99$</td>
</tr>
<tr>
<td>AT3G26590</td>
<td>iMEG</td>
<td>7*, 15, 16, 18*, 19, 20*, 372, 491, 492*, 494*, 495, 496*, 498, 499*</td>
<td>$14 &gt; 0.5$, $7 &gt; 0.95$, $5 &gt; 0.99$</td>
</tr>
<tr>
<td>AT4G26140</td>
<td>iMEG</td>
<td>2, 4, 5*, 6, 7, 9, 10, 497, 518*, 553, 599, 630*</td>
<td>$12 &gt; 0.5$, $3 &gt; 0.95$, $1 &gt; 0.99$</td>
</tr>
<tr>
<td>Genbank ID</td>
<td>Domain</td>
<td>Sites Selected</td>
<td>Posterior Probabilities</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>AT1G55050</td>
<td>iMEG</td>
<td>83*, 84*, 93, 98, 103, 112, 396, 453, 468, 473, 474, 477, 481, 484, 488*, 517, 518, 519*, 520, 525*, 527, 549*, 625, 626, 630, 631, 658, 660, 661, 672, 673, 676, 703, 704, 748, 750, 752, 753, 754, 755*, 756*, 758, 759*, 764, 765, 767*, 769*, 771, 836, 838, 840, 843, 847, 858, 860, 865, 866, 867, 868, 870</td>
<td>60 &gt; 0.5, 11 &gt; 0.95, 1 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G23320</td>
<td>iPEG</td>
<td>5, 8*, 11*, 12, 14, 15*, 16, 17, 18, 20*, 382, 383*</td>
<td>12 &gt; 0.5, 5 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT4G20800</td>
<td>iPEG</td>
<td>4, 5, 20*, 294, 295*, 317, 363, 434, 436*, 493, 526, 527</td>
<td>12 &gt; 0.5, 3 &gt; 0.95, 1 &gt; 0.99</td>
</tr>
<tr>
<td>AT1G17770</td>
<td>iPEG</td>
<td>3, 4, 5, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 23, 33, 37, 60, 61, 79, 83, 88, 91, 93, 98, 101, 104, 105, 128, 141, 144, 145, 146, 160, 162, 202, 204, 210, 252, 334, 389, 411, 413, 447, 461, 467, 472, 475, 481, 501, 506, 509, 510, 527, 552, 574, 596, 667, 672</td>
<td>58 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.99</td>
</tr>
<tr>
<td>AT5G54650</td>
<td>iMEG</td>
<td>14, 57, 176, 177, 189, 191, 202, 313, 323, 336, 363, 432, 492</td>
<td>13 &gt; 0.5, 0 &gt; 0.95, 0 &gt; 0.95</td>
</tr>
</tbody>
</table>

- Positively selected sites with posterior probability > 0.95 are marked with asterisk (*)
5.3.4 Positive selection in imprinted genes compared to random set selected from genome and endosperm

Further, I have investigated whether *A. thaliana* imprinted genes have experienced elevated level of positive selection during their evolution compared to the rest of the genes in *A. thaliana*. Due to the lack of whole genome estimates of positive selection in *A. thaliana*, a random set of 100 genes was selected from *A. thaliana* genome. As imprinting in plants is mainly observed in endosperm, thereby along with the random set from the whole genome, a random set of 100 genes was selected from the identified endosperm-specific genes (Belmonte et al., 2013). Results of the conducted analysis demonstrate that the iPEGs have shown higher percentage of positively selected genes for both lineage-specific selection and site-specific selection compared to the random datasets (Figure 5.5 and Figure 5.6).

To test the statistical significance of this higher percentage of positive selection observed in iPEGs, a fisher-test was performed to test: a) whether iPEGs or iMEGs are experiencing more positive selection compared to random set from genome b) if iPEGs or iMEGs are experiencing more positive selection compared to a random set selected from endosperm-specific genes. Interestingly, iPEGs have experienced elevated levels of lineage-specific positive selection compared to both genome (fisher test, P= 0.0007) and endosperm random datasets (fisher test, P= 0.0034). For iPEGs, the odds ratio score for lineage-specific positive selection indicated 4.2- and 3.3-fold enrichment in positive selection in imprinted genes compared to whole-genome and endosperm controls, respectively. However, no significant differences were observed for site-specific selection in iPEGs compared to both the random datasets (fisher test, P= 0.0758 for genome random set and fisher test, P= 0.2272 for endosperm random set). iMEGs have not shown any enrichment for positive selection both for lineage-specific (fisher test, P= 0.2244 for genome random set and fisher test, P= 0.3717 for endosperm random set) and site-specific selection (fisher test, P=0.3181 for genome random set and fisher test, P= 0.5614 for endosperm random set).
Figure 5.5: Represents the percentage of lineage-specific positively selected genes identified in imprinted genes and random sets. X-axis represents iMEGs, iPEGs and random sets (endosperm and genome), % of genes subject to imprinting are shown on Y-axis.
Figure 5.6: Represents the percentage of site-specific positively selected genes identified in imprinted genes and random sets. X-axis represents iMEGs, iPEGs and random sets (endosperm and genome), % of genes subject to imprinting are shown on Y-axis.
5.3.5 Most of the imprinted genes displayed fixation of positively selected sites (posterior probability > 0.95) within *A. thaliana* population

Positive selection may also lead to fixation of beneficial mutations within a population (Sabeti et al., 2006). Therefore, the percentage conservation of *A. thaliana*-specific (or lineage-specific) positively selected sites was identified. Most of the imprinted genes containing *A. thaliana*-specific positive selected sites (posterior probability > 0.95) displayed fixation of sites (100% conservation) in *A. thaliana* 80 accessions data (Table 5.5). Only two imprinted genes, AT1G48910 and AT1G55050, have shown sites with variable codons. Data for percentage of conservation for all the positively selected sites (posterior probability > 0.5) is given in Supplementary Table 5.6.

Table 5.5: Percentage of fixation of lineage-specific positively selected sites (posterior probability > 0.95) in Ath population data.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Sites</th>
<th>Posterior probability</th>
<th>Ath amino acid</th>
<th>% of conservation of sites in Ath 80 accessions data</th>
<th>Aly amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT2G28380</td>
<td>272</td>
<td>0.982</td>
<td>N</td>
<td>100</td>
<td>A</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>108</td>
<td>0.951</td>
<td>D</td>
<td>100</td>
<td>K</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>178</td>
<td>0.962</td>
<td>E</td>
<td>100</td>
<td>N</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>225</td>
<td>0.974</td>
<td>E</td>
<td>100</td>
<td>P</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>581</td>
<td>0.966</td>
<td>L</td>
<td>100</td>
<td>A</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>240</td>
<td>0.954</td>
<td>L</td>
<td>100</td>
<td>NA</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>342</td>
<td>0.978</td>
<td>A</td>
<td>100</td>
<td>NA</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>441</td>
<td>0.95</td>
<td>L</td>
<td>100</td>
<td>NA</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>1240</td>
<td>0.975</td>
<td>S</td>
<td>100</td>
<td>F</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>153</td>
<td>0.992</td>
<td>G</td>
<td>96.25</td>
<td>P</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>60</td>
<td>0.994</td>
<td>I</td>
<td>100</td>
<td>P</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>325</td>
<td>0.996</td>
<td>N</td>
<td>100</td>
<td>S</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>357</td>
<td>0.992</td>
<td>V</td>
<td>100</td>
<td>N</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>285</td>
<td>0.99</td>
<td>E</td>
<td>100</td>
<td>P</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>381</td>
<td>0.975</td>
<td>A</td>
<td>100</td>
<td>K</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>179</td>
<td>0.955</td>
<td>K</td>
<td>100</td>
<td>R</td>
</tr>
<tr>
<td>AT1G20190</td>
<td>308</td>
<td>0.996</td>
<td>V</td>
<td>100</td>
<td>P</td>
</tr>
<tr>
<td>AT1G20190</td>
<td>195</td>
<td>0.995</td>
<td>C</td>
<td>100</td>
<td>H</td>
</tr>
<tr>
<td>AT1G20190</td>
<td>179</td>
<td>0.973</td>
<td>Y</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>AT1G20190</td>
<td>97</td>
<td>0.969</td>
<td>A</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Amino Acid</td>
<td>Score</td>
<td>Frequency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>-------</td>
<td>-----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>N</td>
<td>0.969</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>0.969</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>0.962</td>
<td>100</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>I</td>
<td>0.997</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>K</td>
<td>0.996</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>R</td>
<td>0.994</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>S</td>
<td>0.993</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>P</td>
<td>0.993</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>K</td>
<td>0.992</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>K</td>
<td>0.991</td>
<td>100</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>Q</td>
<td>0.989</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>N</td>
<td>0.984</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>G</td>
<td>0.984</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>V</td>
<td>0.981</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>D</td>
<td>0.976</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>T</td>
<td>0.971</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>K</td>
<td>0.963</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>R</td>
<td>0.961</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>AT4G31900</td>
<td>K</td>
<td>0.96</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT1G54280</td>
<td>C</td>
<td>0.971</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>AT1G49600</td>
<td>D</td>
<td>0.99</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>AT1G67820</td>
<td>P</td>
<td>0.982</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>AT1G67820</td>
<td>S</td>
<td>0.975</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>S</td>
<td>1</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>K</td>
<td>0.992</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>K</td>
<td>0.988</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>R</td>
<td>0.986</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>E</td>
<td>0.982</td>
<td>100</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>K</td>
<td>0.979</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>F</td>
<td>0.978</td>
<td>92.5</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>A</td>
<td>0.976</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>P</td>
<td>0.974</td>
<td>100</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>D</td>
<td>0.973</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>Y</td>
<td>0.971</td>
<td>85</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>T</td>
<td>0.961</td>
<td>98.75</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>G</td>
<td>0.957</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>L</td>
<td>0.957</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>S</td>
<td>0.956</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>E</td>
<td>0.953</td>
<td>83.75</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>R</td>
<td>0.952</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>AT1G55050</td>
<td>E</td>
<td>0.951</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>AT2G32370</td>
<td>Y</td>
<td>0.976</td>
<td>100</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>AT2G32370</td>
<td>S</td>
<td>0.974</td>
<td>100</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>AT2G32370</td>
<td>N</td>
<td>0.961</td>
<td>100</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
* Genes with no *A. lyrata* ortholog present in the tree are marked with NA. *A. lyrata* ortholog is either not present or discarded from alignments due to presence of recombination.

Further I have tried to assess the functional importance of the predicted *A. thaliana*-specific positively selected sites by testing if these sites are overlapping with any known domains, regions or functional site. Out of 21 genes with *A. thaliana*-specific positively selected sites, 12 genes were annotated with known functional sites, domains or regions in UniProt database. Our lab has conducted wet-lab experiments to identify the functional relevance of these genes (Peter Ryder and Peter McKeown, unpublished data). Fertility rate (number of seeds) and percentage of A (seed abortion), N (growing seeds) and U (unfertilized ovules) was calculated in the knockout of these genes. A significant increase in the number of unfertilized ovules was observed for AT1G61330, which could be linked to *A. thaliana*-specific positive selection identified. Further work will be carried out to confirm if the phenotypic changes identified are due to one of the positive selected sites identified. I discuss the positive selection results in detail for AT1G61330.

AT1G61330 is an FBD-associated F-box protein. The orthology cluster for this gene contains 42 sequences that represent 20 plant species. In *A. thaliana* 0.89% sites were identified to be under positive selection for this gene. All the sites identified showed 100% conservation in *A. thaliana* 80 accession population data (Table 5.5). Positively selected sites 15 and 32 were identified to be in F-box domain of this gene (Figure 5.7).

**Table 5.6: Percentage of conservation of lineage-specific positively selected sites in Ath 80 population data identified in AT1G61330.**

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Sites</th>
<th>Posterior probability</th>
<th>Ath amino acid</th>
<th>% of conservation of sites in Ath 80 accessions</th>
<th>Aly amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G61330</td>
<td>15</td>
<td>0.898</td>
<td>L</td>
<td>100</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>0.946</td>
<td>T</td>
<td>100</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.567</td>
<td>K</td>
<td>100</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>429</td>
<td>0.589</td>
<td>V</td>
<td>100</td>
<td>Y</td>
</tr>
</tbody>
</table>
Figure 5.7: Positively selected sites identified for AT1G61330. The $X$-axis represents the start and the end of the gene. The $Y$-axis represents the posterior probability. Blue regions represent the domains. Green triangles represent the *A. thaliana*-specific positive selected sites.
5.3.6 Both iPEGs and iMEGs showed an excess of deleterious mutations in *A. thaliana* population data

No evidence of positive selection is identified in imprinted genes with ≤ 5 species using McDonald-Kreitman test (64 imprinted genes, where all-vs-all BLASTP identified *A. lyrata* orthologs for 49 imprinted genes). Further I performed the McDonald-Kreitman test on the full set of 148 imprinted genes to identify the rate of evolution of imprinted genes within the *A. thaliana* population. All-vs-all BLASTP identified *A. lyrata* orthologs for 118 *A. thaliana* imprinted genes (58 iPEGs and 60 iMEGs). The $P_N/P_S$ ratio obtained for *A. thaliana* populations was 1.196 (Table 5.7), which is higher than the $D_N/D_S$ ratio for fixed differences in the same set of genes (1.139). An eFI (Axelsson and Ellegren, 2009) 1.205 was obtained for imprinted genes, which is greater than $FI$ 0.952 obtained. The negative value of $\alpha$ (-0.210) for imprinted genes indicates an excess of slightly deleterious mutations in *A. thaliana* populations $\alpha=(0.952-1.205)/1.205$). Value of $\alpha$ is obtained negative for both iPEGs and iMEGs, however iPEGs showed higher value of $\alpha$ (-0.198) compared to iMEGs (-0.225).

The McDonald-Kreitman test identified only one imprinted gene (AT1G17770) with significantly higher $D_N/D_S$ than $P_N/P_S$ (NI 0.479, $\alpha$ 0.52 and p < 0.05) (Supplementary Table 5.5). Interestingly, this gene is identified to be under lineage-specific selection in 34 sequenced species data using codeML models. An additional 13 imprinted genes were identified with significantly higher $P_N/P_S$ than $D_N/D_S$ (p < 0.05).

Table 5.7: Number of non-synonymous and synonymous substitutions, Observed $FI$ and e$FI$ for *A. thaliana* imprinted genes.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Polymorphism</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-synonymous substitutions</td>
<td>1988</td>
<td>4740</td>
</tr>
<tr>
<td>Synonymous substitutions</td>
<td>1662</td>
<td>4161</td>
</tr>
<tr>
<td>Non-synonymous /synonymous</td>
<td>1.196</td>
<td>1.139</td>
</tr>
<tr>
<td>Fixation index ($FI$)</td>
<td></td>
<td>0.952</td>
</tr>
<tr>
<td>Expected fixation index (e$FI$)</td>
<td></td>
<td>1.205</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td>-0.210</td>
</tr>
</tbody>
</table>
5.3.7 Imprinted genes with disrupted reading frame in *A. thaliana* accessions showed high *Dn/Ds* and high *Pn/Ps* ratios

To identify if *A. thaliana* imprinted genes show any clustering in the 2D plot of *Dn/Ds* and *Pn/Ps* ratios, I have conducted Mcdonald-Kreitman analysis on all the protein-coding genes of *A. thaliana* (Figure 5.8). 2D plot of *Dn/Ds* and *Pn/Ps* ratios is the decomposition of classical Mcdonald-Kreitman test (Daub et al., 2014). The positions in the 2D plot represents the following potential evolutionary forces: 1) low *Dn/Ds* and low *Pn/Ps* represents ancient strong purifying selection, 2) high *Dn/Ds* and low *Pn/Ps* represents ancient positive selection or recent purifying selection, 3) high *Pn/Ps* and low *Dn/Ds* represents balancing selection or recent positive selection, 4) high *Pn/Ps* and high *Dn/Ds* represents ancient relaxation or pseudogenization.

No major dominance of *A. thaliana* imprinted genes was observed in any of the four categories (Figure 5.8). Interestingly, six of the genes with high *Dn/Ds* and high *Pn/Ps* ratios were identified to be pseudogenes in *A. thaliana* (Yang et al., 2011; Wang et al., 2012) (Table 5.8). We have not identified any evidence of lineage-specific or site-specific selection for these six pseudogenes genes. Out of six pseudogenes genes, four were found to be uncharacterized. These results suggest that accelerated evolutionary rates might serve as a proxy for identification of pseudogenes.

**Table 5.8: High *Dn/Ds* and High *Pn/Ps* of *A. thaliana* imprinted genes identified as pseudogenes.**

<table>
<thead>
<tr>
<th>GeneId</th>
<th><em>Dn/Ds</em></th>
<th><em>Pn/Ps</em></th>
<th>Mutation type</th>
<th>Number of accessions with broken gene models</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G29570</td>
<td>5.46</td>
<td>2.75</td>
<td>PREMATURE</td>
<td>1</td>
</tr>
<tr>
<td>AT3G57250</td>
<td>2.30</td>
<td>5.50</td>
<td>FRAMESHIFT</td>
<td>14</td>
</tr>
<tr>
<td>AT1G11810</td>
<td>2.91</td>
<td>5.00</td>
<td>FRAMESHIFT</td>
<td>31</td>
</tr>
<tr>
<td>AT1G61090</td>
<td>4.33</td>
<td>2.00</td>
<td>FRAMESHIFT</td>
<td>3</td>
</tr>
<tr>
<td>AT1G07690</td>
<td>2.58</td>
<td>4.00</td>
<td>FRAMESHIFT</td>
<td>1</td>
</tr>
<tr>
<td>AT1G20730</td>
<td>3.88</td>
<td>5.50</td>
<td>FRAMESHIFT</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 5.8: 2D plot of $D_n/D_s$ and $P_n/P_s$ of *A. thaliana* protein-coding genes. X-axis depicts $P_n/P_s$ ratios. Y-Axis represents $D_n/D_s$ ratios. Green dots- genes under purifying selection, red dots- genes under positive selection, yellow dots- genes under neutral evolution and black triangles- *A. thaliana* imprinted genes, blue triangles- pseudogenes with high $D_n/D_s$ and high $P_n/P_s$. 
5.4 Discussion

This study was conducted to identify signatures of positive selection on experimentally validated set of 148 A. thaliana imprinted genes. Out of 148 genes, nearly half of the set (74 genes) showed very less number of species (< 6) in their orthology clusters (Figure 5.2). Interestingly, orthology clusters data for imprinted genes have shown that almost all the imprinted genes belong to multigene families (only one single-copy ortholog gene AT4G31060). This data indicated that imprinted clade for A. thaliana imprinted genes might have recently evolved. This hypothesis is further supported by the enrichment of imprinted genes for recent whole genome duplication (alpha events) that are specific to Brassicaceae family (section 5.3.2).

Previous studies have reported signatures of positive selection on imprinted genes, like MEDEA (Spillane et al., 2007). Positive selection on imprinted genes has been understood as a signature of conflict between maternal and paternal genomes over allocation of resources. However, the study conducted on mammalian imprinted genes did not show any correlation between the imprinted status and evidence of positive selection (O’Connell et al., 2010). Interestingly, the analysis conducted in this chapter demonstrated a significant elevated level of A. thaliana-specific positive selection in iPEGs compared to random dataset collected from genome and endosperm (Figure 5.5). Whereas, iMEGs did not show any significant difference compared to both random sets. Both iMEGs and iPEGs did not show any evidence of site-specific positive selection (Figure 5.6). One possible explanation for high rate of positive selection in iPEGs could be the sexual conflict between male gametophyte-expressed genes. Gossmann et al. (2014) suggested that pollen tube competition allows pre-fertilization sexual conflict between male genomes to occur in flowering plants in an analogous manner to the competition between mammalian sperm (Gossmann et al., 2014). This sexual conflict suggests that positive selection could also be triggered by conflict between the paternal genomes of endosperm in different ovules developing on the same plant. If this is the case than high rate of positive selection in iPEGs is in agreement with this hypothesis.

Unlike mammalian imprinted genes, imprinting status of A. thaliana genes is not known in other plant species. Therefore, I could not use the strategy by O’Connell et
al. (2010) to demonstrate the correlation between imprinting status and evidence of positive selection in multiple species (O’Connell et al., 2010).

Gene age is a strong predictor of the rate of molecular evolution (Cai and Petrov, 2010). Gene age analysis demonstrated no significant age difference between iMEGs and iPEGs. Therefore, the evidence of positive selection identified is not an artifact of difference in gene age. Positive selection may also lead to fixation of beneficial mutations in a population (Sabeti et al., 2006). A fixation (100% conservation) is observed for most of the imprinted genes containing *A. thaliana*-specific positive selected sites (Table 5.5). The 12 imprinted genes identified here (with positive selected sites probability >0.95 and fixation in *A. thaliana* population data) are strong candidates for further functional studies.

Wolff et al. (2011) suggested the possibility that imprinted genes are on a trajectory to become pseudogenes. As high *Dn/Ds* and high *Pn/Ps* suggest the possibility of ancient relaxation or pseudogenization (Daub et al., 2014), I have conducted genome-wide analysis to identify whether imprinted genes show any clustering in 2D plot of *Dn/Ds* and *Pn/Ps* ratios. However, no clustering for high *Dn/Ds* and high *Pn/Ps* (which is indicative of pseudogenization) is observed for *A. thaliana* imprinted genes (Figure 5.8) or enrichment in pseudogenes identified in *A. thaliana* (Yang et al., 2011; Wang et al., 2012).
5.5 Conclusions and future work

Results in this chapter have demonstrated a significant elevated level of *A. thaliana*-specific positive selection in *A. thaliana* iPEGs. Most of these genes have shown fixation of positively selected sites in *A. thaliana* population data, which will be our candidates for further functional studies. Genome-wide 2D decomposition of McDonald-Kreitman test conducted has not identified any signatures of imprinted genes becoming non-functional pseudogenes.
Chapter 6. Detection of imprinting and dosage effects in triploid *Arabidopsis thaliana* embryos
6.1 Introduction

Genomic imprinting refers to a phenomenon independent of Mendelian genetics where only one allele is actively expressed depending on its parent of origin. This phenomenon affects a subset of genes in flowering plants and mammals (Garnier et al., 2008; Kohler et al., 2012; McKeown et al., 2013). Imprinting in plants is mainly observed in endosperm, a triploid nutritive tissue (2 maternal: 1 paternal genome complement) consumed by the embryo during seed development or germination. Although considered rare, a few cases of imprinting are also observed in embryo (Jahnke and Scholten, 2009; McKeown et al., 2013). Diploid genome embryo nuclei contain 1 maternal: 1 paternal genome complement. Nodine and Bartel (2012) suggested that imprinting may occur in the A. thaliana embryo during very early stages of development (Nodine and Bartel, 2012).

Imprinting can affect all/most of the alleles (gene-specific imprinting) or specific alleles of a gene (allele-specific imprinting) (Baroux et al., 2002). For example, allele-specific imprinting was reported for the PHERES1 (PHE1) gene in Arabidopsis thaliana, expression of which is paternal in C24 X Col-0 crosses but shows biallelic expression in the opposite (Col-0 X C24) cross direction (Villar et al., 2009).

Although the hypotheses to explain the evolution of imprinting are under constant debate (Haig, 2013), the kinship hypothesis (or parental conflict) is the most tested and widely supported theory (Hurst and McVean, 1998; Wilkins and Haig, 2003). This theory states that the imprinting has evolved because of an evolutionary conflict in individuals between maternally and paternally derived alleles over maternal resource allocation and optimal offspring size. This conflict can cause growth-promoting genes to be active when inherited paternally but silenced when inherited maternally (Haig and Westoby, 1989). The phenomenon of imprinting has been demonstrated to be disrupted by interploidy crosses (Erilova et al., 2009; Jullien and Berger, 2010; Tiwari et al., 2010), where increasing the dosage of silenced allele of an imprinted gene can lead to removal of the imprinting marks and reactivation of the allele. This has led to an alternative hypothesis that the triploid genome in the endosperm might be required for imprinting in plants.
Interploidy crosses between diploid and tetraploid Arabidopsis thaliana leads to the formation of triploid F1 progeny. It is possible in reciprocal crosses F1 triploid system to answer whether: 1) genomic imprinting is an effect of the triploid genome dosage? 2) imprinting can be artificially created in tissues other than endosperm which are normally diploid, and where imprinting is not normally observed? Donoghue et al. (2013) demonstrated the evidence of widespread parent-of-origin-specific genome-dosage effects on gene expression in isogenic reciprocal F1 triploids that lead to the differential expression of 602 genes in Col-0 F1 triploids (Donoghue et al., 2013). Hybrid triploid data can be further used to identify whether the gene dysregulation observed in isogenic triploids is due to differential parental allele expression. In F1 hybrids, any parental bias in expression can be identified based on the SNPs (Single Nucleotide Polymorphism) present in the transcribed regions between two parent accessions.

To address these questions, our lab (unpublished data) has generated a number of F1 triploid hybrid embryos between Col-0 and C24. Considering either Col-0 or C24 as tetraploid parents, an imprinting-like phenomenon was observed on a set of genes where in paternal excess triploids, the maternal alleles are up-regulated, compared to the paternal alleles of maternal excess triploids (experimental work conducted by Antoine Fort). This phenomenon is observed independent of the accession (Col-0 or C24) used as a tetraploid or diploid parent. Figure 6.1 and 6.2 demonstrates the high maternal allele contribution for three genes in paternal excess triploid embryos dissected out of their developing seeds. To identify whether this phenomena occurs on a genome-wide basis, our lab has generated RNA-seq data for maternal excess triploid (4xC24 X 2xCol) and paternal excess triploid (2xCol X 4xC24) embryos where we want to study allele-specific expression on a genome-wide basis. C24 is used as a tetraploid parent for whole transcriptome analysis, as using Col-0 as a tetraploid parent leads to high F1 seed abortion and developmental abnormalities.
Figure 6.1: Allelic contributions of three genes in reciprocal triploid hybrids using Col-0 as a tetraploid parent (experimental work conducted by Antoine Fort).

Figure 6.2: Allelic contributions of three genes in reciprocal triploid hybrids using C24 as a tetraploid parent (experimental work conducted by Antoine Fort).
6.2 Materials and Methods

6.2.1 High-throughput sequencing and data filtering

Short-read paired-end data (read size 100 bp and insert-size 150 bp) were generated for Col-0 and C24 F1 triploid hybrid embryos (4xC24 X 2xCol and 2xCol X 4xC24), each sample with three biological replicates. Low-quality bases from both the 5’- and 3’- end of the reads were trimmed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). Bases with error rate > 1% (Phred score <20) were considered as low-quality. Reads shorter than 25 bp after trimming were also discarded from further analysis.

6.2.2 SNP data

We have downloaded the already identified SNP data between Col-0 and C24 from 1001 genome website (Schneeberger et al., 2011). SNP coordinates were translated from TAIR8 to TAIR10 using the translate_tair8.pl script provided on 1001 genome website (http://1001genomes.org/data/software/translate_tair8).

6.2.3 Mapping reads on Col-0 and C24 genomes

The major technical challenge with identification of allele-specific expression is reference bias that can be defined as the effect where the reads possessing the reference allele mapped better than the reads carrying the non-reference allele (Wang and Clark, 2014). To overcome this problem we have mapped short read data on both Col-0 and C24 genomes. Allele-specific expression (ASE) can be measured by RNA-seq by assigning reads to the individual alleles. Genes with expression deviation from the expected ratio can be classified as either having larger paternal expression or maternal expression.

Paired-end reads for both the maternal excess and paternal excess samples were mapped to both Col-0 (TAIR10) and C24 genomes using TopHat 1.4.1 (Trapnell et al., 2009). C24 genome was downloaded from the 1001 genome website. Read mapping was done using the default mapping criteria in Tophat (and --mate-inner-dist -50 bp as
per our dataset). Samtools package (Li et al., 2009) were used to do the downstream analysis. A custom made python script (utilizing pysam python module) was used to calculate the number of reads representing reference and non-reference SNP. If a gene contained multiple SNPs then the grand total of all the reads containing reference and non-reference SNPs were considered as the expression frequency of that gene. Only genes with total expression frequency greater than 20 reads were considered for further analysis. A SNP with read depth less than 10 was discarded.

6.2.4 Determining Allele-specific expression

For determining the allele-specific expression, we calculated the probability of a gene’s expression deviation from expectation using binomial tests. The expected SNP ratio present between the alleles should be 66% for C24 and 33% for Col-0 (for triploid tissues made from a C24 tetraploid and Col-0 diploid). For each gene and sample (maternal excess and paternal excess), we performed a binomial one-tailed test under the null hypotheses 2m:1p in case of maternal excess (4xC24 X 2xCol) and 1m:2p in case of paternal excess (2xCol X 4xC24). Multiple correction was done using q-values (Dabney and Storey, 2015). Genes with q-value less than 0.05 were identified as genes with expression deviated from the expected ratio towards either larger paternal expression or maternal expression.

In addition to imprinting effects, biased expression patterns can also be a result of accession-specific effects on gene expression. For example, the Col-0 allele of a particular gene might be more highly expressed than a C24 allele, independent of the direction of the cross. A gene was considered as dominant if it showed significant expression deviation for one accession in both the triploid hybrids. We called the candidate imprinted genes based on significant deviation in expression and the degree of allele-specific bias (as suggested by (Wang and Clark, 2014)), using an arbitrary cut-off of expression ratio > 75% for C24 and > 45% for Col-0. For example, a gene is considered to be maternally expressed if 1) it showed significant bias maternal bias to express the maternal allele in both maternal excess and paternal excess triploids, 2) showed expression ratio > 75% for C24 in maternal excess triploids and > 45% for Col-0 in paternal excess triploids.
6.3 Results

6.3.1 A significantly higher number of genes showed maternal bias in both the maternal excess and paternal excess triploids

Parental allele bias in maternal excess triploids (4xC24 X 2xCol) and paternal excess triploids (2xCol X 4xC24) embryos was assessed at whole transcriptome level. Genes with expression deviation from the expected SNP ratios (66% for C24 and 33% for Col-0) were identified as genes with biased allelic expression. A total of 8,006 genes did not show any expression deviation from the expected SNP ratios in both the F1 triploids and were considered as unbiased genes. A significant deviation in expression was observed for 3,838 (29.2%) and 3,917 (29.3%) genes in maternal excess and paternal excess triploids respectively. Apart from imprinting effects, biased gene expression can also be a result of accession-specific effects on gene expression, also termed dominant genes (Meyer et al., 2007). A total of 2,603 genes showed dominant expression either from Col-0 or C24 (Figure 6.3). A significant higher number of dominant genes were observed in Col-0 (1,616 genes) compared to C24 (987 genes) (Fisher test $P < 2.2e-16$).

After filtering dominant genes, a total of 887 genes (6.63%) were identified with maternally biased expression and 313 genes (2.34%) with paternally biased gene expression in paternal excess triploids (2xCol X 4xC24). In maternal excess triploids (4xC24 X 2xCol), 740 genes (5.62%) showed maternally biased gene expression and 440 (3.34%) genes paternally biased gene expression (Figure 6.4). In both the reciprocal crosses, a significantly higher number of maternally biased genes were observed compared to paternally biased genes (Binomial one-tailed test; paternal excess triploids $P < 2.2e-16$, maternal excess triploids $P < 2.2e-16$).
Figure 6.3: Percentage of genes showed dominant gene expression from Col-0 and C24. X-axis represents Col-0 and C24 accessions, % of dominant genes in each accessions is shown on Y-axis.
Figure 6.4: Percentage of genes that showed paternal or maternal bias in maternal excess triploids (4xC24 X 2xCol) and paternal excess triploids (2xCol X 4xC24). Genes with expression deviation from expected SNP ratio (66% for C24 and 33% for Col-0) are categorized as maternal or paternal bias in triploid hybrids.
6.3.2 Gene-specific imprinting observed in reciprocal F1 triploid hybrid embryos

Genomic imprinting in plants is observed to be restricted to endosperm, with a very few exceptions in embryo (Jahnke and Scholten, 2009; Raissig et al., 2013). A cutoff of expression ratio > 75% for C24 and > 45% for Col-0 was used to identify candidate maternally expressed (MEGs) and paternally expressed (PEGs) imprinted genes. A total of 11 candidate gene-specific imprinted genes are identified in reciprocal F1 triploid hybrid embryos (Table 6.1). Figure 6.5 shows the distribution of allelic contribution of unbiased genes, dominant genes, allele-specific imprinted genes and gene-specific imprinted genes in transcriptomes of maternal excess and paternal excess triploids.

Table 6.1: Genes identified to be imprinted (MEGs and PEGs) in maternal and paternal excess triploid data.

<table>
<thead>
<tr>
<th>Gene Id</th>
<th>Imprinting Status</th>
<th>Refallele (Col-0) Percentage in 4xC24 X 2xC0l</th>
<th>Refallele (Col-0) Percentage in 2xC0l X 4xC24</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G14520</td>
<td>Maternal</td>
<td>1.163</td>
<td>94.32</td>
</tr>
<tr>
<td>AT4G12960</td>
<td>Maternal</td>
<td>0.860</td>
<td>100.0</td>
</tr>
<tr>
<td>AT1G50650</td>
<td>Maternal</td>
<td>13.85</td>
<td>48.04</td>
</tr>
<tr>
<td>AT5G59310</td>
<td>Maternal</td>
<td>0.239</td>
<td>100.0</td>
</tr>
<tr>
<td>AT5G47150</td>
<td>Maternal</td>
<td>15.39</td>
<td>63.51</td>
</tr>
<tr>
<td>AT2G38900</td>
<td>Maternal</td>
<td>3.704</td>
<td>72.00</td>
</tr>
<tr>
<td>AT3G20210</td>
<td>Maternal</td>
<td>19.92</td>
<td>75.24</td>
</tr>
<tr>
<td>AT1G19640</td>
<td>Maternal</td>
<td>8.929</td>
<td>100.0</td>
</tr>
<tr>
<td>AT4G03610</td>
<td>Paternal</td>
<td>49.40</td>
<td>19.42</td>
</tr>
<tr>
<td>AT4G13494</td>
<td>Paternal</td>
<td>52.25</td>
<td>15.39</td>
</tr>
<tr>
<td>AT4G13495</td>
<td>Paternal</td>
<td>62.28</td>
<td>20.04</td>
</tr>
</tbody>
</table>
Figure 6.5: Distribution of allelic contribution of biased and unbiased genes in transcriptomes of paternal excess and maternal excess triploids. Genes that showed biased gene expression of one allele (Col-0 or C24) depending on whether the allele is transmitted maternally or paternally in reciprocal triploids, were considered as allele-specific imprinted genes. If a parent-of-origin effect was observed in both cross directions for each allele, these were considered as gene-specific imprinted genes. Genes that showed biased gene expression for one allele in both the reciprocal triploids were considered as dominant genes.
6.3.3 Overlap of candidate imprinted genes identified in F1 triploid embryos with the known imprinted genes in *A. thaliana* endosperm

As described in Chapter 5, a validated set of 148 *A. thaliana* imprinted genes were compiled from various high-throughput screens (Gehring et al., 2011; Hsieh et al., 2011; McKeown et al., 2011; Shirzadi et al., 2011; Wolff et al., 2011b) and other studies (Vielle-Calzada et al., 1999; Kinoshita et al., 2004; Kohler et al., 2005; Jullien et al., 2006; Tiwari et al., 2008; Gehring et al., 2009) (Supplementary Table 5.1). Candidate gene-specific and allele-specific imprinted genes in triploid embryos were compared with the validated set of known imprinted genes. Out of 11 embryo specific imprinted genes, one gene (AT4G13495) was previously known to be imprinted in endosperm (Gehring et al., 2011). A total of four allele-specific imprinted genes in paternal excess triploids and three allele-specific imprinted genes in maternal excess triploids overlapped with the validated set of imprinted genes (Table 6.2).

Table 6.2: Overlap of candidate imprinted genes identified in triploid embryo with the known imprinted genes in *A. thaliana* endosperm.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Imprinting in F1 triploid embryos</th>
<th>Imprinting status in endosperm</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G13495</td>
<td>Gene-specific</td>
<td>iPEG</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G10950</td>
<td>Paternal excess</td>
<td>iPEG</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G05440</td>
<td>Maternal excess</td>
<td>iPEG</td>
<td>McKeown et al., 2011</td>
</tr>
<tr>
<td>AT4G13460</td>
<td>Paternal excess</td>
<td>iPEG</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G49600</td>
<td>Paternal excess</td>
<td>iPEG</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G54650</td>
<td>Maternal excess</td>
<td>iMEG</td>
<td>Gerald et al., 2009</td>
</tr>
<tr>
<td>AT5G22920</td>
<td>Maternal excess</td>
<td>iMEG</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G62340</td>
<td>Paternal excess</td>
<td>iMEG</td>
<td>Wolff et al., 2011</td>
</tr>
</tbody>
</table>
6.4 Discussion

In animal species, maternal gene products deposited into eggs regulate embryonic development before activation of the zygotic genome (Tadros and Lipshitz, 2009). However in plants, uniparental expression patterns in the developing F1 seed could occur via the maternal deposition of mRNAs in the central cell or egg cell, or via paternal deposition of such mRNAs in the central cell or egg cell (reviewed in (McKeown et al., 2013)). For example, *Short Suspensor (SSP)* gene in *A. thaliana* is highly expressed in pollen but not translated. Instead, the *SSP* transcripts are delivered via the sperm cells to the zygote and the endosperm, where *SSP* protein transiently accumulates (Bayer et al., 2009).

Recent genome wide studies conducted to define the allelic contribution in diploid plant embryos have shown contradictory results. Autran et al. (2011) showed an overrepresentation of maternal transcripts in early stage embryos, whereas Nodine and Bartel (2012) concluded the biallelic expression of diploid embryo transcripts (Autran et al., 2011; Nodine and Bartel, 2012). These differences in results could be due to endosperm/or seed coat contamination (Fort and Spillane, 2013). To avoid this contamination, the embryos for RNA-seq data were extracted from later stages of seed development. However, Hsieh et al (2011) did not detect any imprinted genes in the diploid embryos at the same stage of seed development (Hsieh et al., 2011).

Genome wide analysis done in reciprocally crossed F1 triploid system in this chapter has shown a number of candidates for gene-specific and allele-specific imprinting which can be induced in embryos. These genes do not overlap with predicted genes regulated by imprinting in early embryo development (Raissig et al., 2013). A significant deviation in expression was observed for ~ 29% of expressed genes in both our triploid hybrids. Excluding the genes that have shown accession specific dominance effects (20% dominant genes), ~ 9% genes showed biased gene expression in paternal excess and maternal excess triploids. These biases in gene expression for one allele (either Col-0 or C24) represent a phenomena equivalent to allele-specific imprinting.

Further, gene-specific imprinting was observed for 11 genes, where parent of origin
specific expression was observed in both cross directions. Altogether, a high number of MEGs (8 MEGs) were identified compared to PEGs (3 PEGs) in F1 triploid embryos. A similar trend is observed for allele-specific imprinted genes where a significant higher number of genes showed maternal bias in both the maternal excess and paternal excess triploids. An overrepresentation of MEGs (80 MEGs) compared to PEGs (68 PEGs) are also reflected in validated set of 148 A. thaliana imprinted genes (Chapter 5; Supplementary Table 5.1). Similarly, MEGs are overrepresented in both rice (Luo et al., 2011) and maize (Guo et al., 2003), where the possibility of false-positive MEGs due to seed coat contamination is very low. The high number of MEGs than PEGs are congruent with the theories based on maternal co-evolution or care (McKeown et al., 2013).

Few candidate imprinted genes in F1 triploid embryos have shown overlap with the known imprinted genes in A. thaliana endosperm (1 gene-specific imprinted, 7 allele-specific imprinted). However, none of imprinting candidates were identified to be imprinted in the diploid embryo dataset (Hsieh et al., 2011). Therefore, the identification of gene specific and allele specific imprinting effects in triploid embryos demonstrates that triploidy can induce imprinting effects, most likely due to the change in genome dosage.
6.5 Conclusions and future plan

The genome wide analysis done in this chapter reveals the identification of novel gene-specific and allele-specific imprinted genes in triploid embryo hybrids. These results demonstrate that altering the genome dosage can induce artificial imprinting in plant tissues where imprinting does not occur at the diploid level. Further experiments will be conducted in diploid and tetraploid hybrids embryos to identify whether such imprinting effects exist in balanced hybrids or are only restricted to unbalanced triploid hybrids. Understanding the parent-of-origin effects is of major importance for the understanding of phenotypic variation in polyploids and crop improvement.
Chapter 7. Summary and future directions
In this thesis it was sought to gain a better understanding of origin and evolution of orphan and imprinted genes in plant species. Orphan genes are important sets of genes that may have played a role in acquiring adaptation to different environmental conditions and other unique characteristics to these species. Therefore, research work done in this thesis on orphan genes in agronomically important crops not only have meaningful evolutionary implications for the fundamental biology of species but can also assist plant breeding. Imprinted genes are another set of important genes that have been argued to facilitate species-specific adaptations. Results presented in the last two chapters of this thesis broaden our understanding on the evolution of plant imprinted genes and critically assess theories concerning their evolution and origins.

Chapter 1, presents literature review on origin and evolution of orphan and imprinted genes. This chapter also includes general information about duplication, selection and mechanism of origination of new genes.

In Chapter 2, orphan genes were identified in the genomes of two crop species, *Camelina sativa* and *Sorghum bicolor*. Both species are important agronomic crop species where *Sorghum bicolor* is the fifth most important crop grown in the world and *Camelina sativa* has potential as a biofuel crop. Apart from their importance as crop species, both species have the potential to adapt to future climate conditions, particularly increasing drought. A total of 1.85% (1,656 genes) and 7.6% (2,499 genes) were identified to be orphan genes in *C. sativa* and *S. bicolor* respectively. In addition to species-specific genes, an additional 2,105 Brassicaceae-restricted orphan genes were identified in *C. sativa*. Further, a Blast-based sequence similarity approach to identify evolutionary origins of orphan genes in both species. In *C. sativa* duplication was identified to be the major mechanism of origination of orphan genes (46%), whereas in *S. bicolor* both de novo (10.6%) and duplication (11.9%) mechanisms displayed equal contribution.

Transcriptomic data of 40 *S. bicolor* accessions was analyzed to identify expression evidence of *S. bicolor* orphan genes. A set of 52 orphan genes were identified as ‘core-set’ of expressed orphan genes in *S. bicolor*. Orphan genes identified in *S. bicolor* will be further analyzed in recently sequenced outgroup species for evidence
of gene models retention. Orphan genes unique to *S. bicolor* will be screened for functionality to identify *S. bicolor* specific adaptations.

In **Chapter 3**, a number of novel chimeric ORFs were identified that display features known to trigger Cytoplasmic-male sterility (CMS) in other plant species. CMS in plants are known to be caused by chimeric ORFs that are produced as a result of mitochondrial genome rearrangements. The work done in this chapter also reports a high-quality pigeonpea mitochondrial genome assembly for male-sterile line ICPA 2039. This work provides the first comparative study of legume mitochondrial genome sequences and identified several rearrangement regions as well as 13 novel candidate chimeric ORFs associated with CMS in pigeonpea.

These chimeric ORFs have been hypothesized to interfere with the normal function of respiratory chain reactions and cause pollen development to abort. The chimeric ORFs identified between male-sterile and fertile line represents the most promising candidates for CMS-related mitochondrial rearrangements in pigeonpea. Of the 13 potential candidates, five carry parts of other mitochondrial genes and eight were observed to be in the proximity to other mitochondrial genes. Interestingly, most of our candidate ORFs carries regions of transmembrane domains, which have been argued to be another prominent feature of CMS ORFs.

Further, structural and functional studies will be conducted to allow the exact mitochondrial genome segments responsible for male-sterility in pigeonpea. Transcription and translation patterns of our candidates will be determined, including in young to mature buds and floral parts including the pollen mother cell, tetrad and pollen grains. The roles of transmembrane domains and respiration in the mitochondrial genome of ICPA 2039 will also be assessed.

In **Chapter 4**, I developed a plant orthology database due to the underrepresentation of plant genomes in existing orthology database at the time of construction of this database. This database has developed as a community resource and for identification of signatures of positive selection on orthologs of *A. thaliana* imprinted genes.
In Chapter 5, I have tested if genomic imprinting in a flowering plant endosperm was associated with positive selection at the loci affected, as predicted by many of the theories for why imprinting occurs. Imprinted genes have been reported as being subject to positive selection in plants and mammals but attempts to link imprinting with positive more generally have been equivocal. In this analysis, an ortholog-based approach was used to demonstrate that in fact positive selection does affect imprinted genes in *A. thaliana*, but that the frequency of positive selection was only significantly elevated at imprinted Paternally Expressed Genes (iPEGs) and not imprinted Maternally Expressed Genes (iMEGs).

A significant elevated level of *A. thaliana*-specific positive selection in iPEGs compared to random dataset collected from genome and endosperm. Whereas, iMEGs did not show any significant difference compared to both random sets. Both iMEGs and iPEGs did not show any evidence of site-specific positive selection. One intriguing result of this analysis of imprinted genes in plants is that differential selective pressures appear to act on imprinted genes expressed from the maternal vs. the paternal genomes: specifically, iPEGs have higher \( d_S/d_S \), and are significantly more likely to be subject to positive selection.

Almost all of the amino acids encoded by nucleotides subject to positive selection were identified to be fixed across populations, indicating that these are likely to have been important for plant fitness and are strong candidates for future functional investigations. Further functional work on genes such as these is likely to yield further examples of the so-far elusive imprinted phenotypes.

In Chapter 6, the genome-wide analysis conducted in reciprocally crossed F1 triploid system reveals the identification of novel gene-species and allele-specific imprinting in triploid embryos hybrids. Studying the imprinting effects is of major importance for the understanding of phenotypic variation in polyploids and crop improvement.

The work done in this chapter demonstrates that artificial imprinting can be induced in plant tissues by altering the genome dosage. These results also support the hypothesis that the triploid genome in the endosperm might be required for imprinting
in plants. Further experimental work will be conducted to generate diploid and tetraploid hybrids embryos between Col-0 and C24. The newly generated diploid and tetraploid dataset will serve as a control set to identify whether the imprinting effects observed in triploid embryos exist in balanced hybrids or are only restricted to unbalanced triploid hybrids.
Supplementary Figure 2.1: Baseline FPKM distribution of 40 sorghum accessions. Genes with FPKM >1 were considered to be expressed (represented with red line).
**Supplementary Table 2.1: De novo originated overprinted genes identified in A. thaliana.**

<table>
<thead>
<tr>
<th>Parent Gene</th>
<th>Gene age</th>
<th>Overprinted Gene</th>
<th>Gene age</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G07490</td>
<td>6</td>
<td>AT1G07485</td>
<td>1</td>
</tr>
<tr>
<td>AT1G08260</td>
<td>9</td>
<td>AT1G08270</td>
<td>9</td>
</tr>
<tr>
<td>AT1G16022</td>
<td>2</td>
<td>AT1G16025</td>
<td>2</td>
</tr>
<tr>
<td>AT1G17745</td>
<td>9</td>
<td>AT1G17744</td>
<td>4</td>
</tr>
<tr>
<td>AT1G19394</td>
<td>0</td>
<td>AT1G19396</td>
<td>0</td>
</tr>
<tr>
<td>AT1G24388</td>
<td>2</td>
<td>AT1G24390</td>
<td>0</td>
</tr>
<tr>
<td>AT1G24880</td>
<td>9</td>
<td>AT1G24822</td>
<td>0</td>
</tr>
<tr>
<td>AT1G26500</td>
<td>7</td>
<td>AT1G26510</td>
<td>2</td>
</tr>
<tr>
<td>AT1G45165</td>
<td>2</td>
<td>AT1G45163</td>
<td>0</td>
</tr>
<tr>
<td>AT1G51400</td>
<td>8</td>
<td>AT1G51402</td>
<td>0</td>
</tr>
<tr>
<td>AT1G62620</td>
<td>9</td>
<td>AT1G62610</td>
<td>9</td>
</tr>
<tr>
<td>AT1G62880</td>
<td>9</td>
<td>AT1G62886</td>
<td>9</td>
</tr>
<tr>
<td>AT1G67856</td>
<td>9</td>
<td>AT1G67855</td>
<td>0</td>
</tr>
<tr>
<td>AT1G68940</td>
<td>8</td>
<td>AT1G68935</td>
<td>0</td>
</tr>
<tr>
<td>AT1G78480</td>
<td>9</td>
<td>AT1G78476</td>
<td>0</td>
</tr>
<tr>
<td>AT1G78700</td>
<td>8</td>
<td>AT1G78710</td>
<td>6</td>
</tr>
<tr>
<td>AT2G04378</td>
<td>2</td>
<td>AT2G04380</td>
<td>0</td>
</tr>
<tr>
<td>AT2G05755</td>
<td>9</td>
<td>AT2G05753</td>
<td>0</td>
</tr>
<tr>
<td>AT2G07827</td>
<td>0</td>
<td>AT2G07830</td>
<td>0</td>
</tr>
<tr>
<td>AT2G16018</td>
<td>2</td>
<td>AT2G16019</td>
<td>0</td>
</tr>
<tr>
<td>AT2G18969</td>
<td>6</td>
<td>AT2G18970</td>
<td>0</td>
</tr>
<tr>
<td>AT2G19340</td>
<td>9</td>
<td>AT2G19330</td>
<td>9</td>
</tr>
<tr>
<td>AT2G24620</td>
<td>6</td>
<td>AT2G24617</td>
<td>0</td>
</tr>
<tr>
<td>AT2G24680</td>
<td>8</td>
<td>AT2G24670</td>
<td>5</td>
</tr>
<tr>
<td>AT2G25410</td>
<td>8</td>
<td>AT2G25409</td>
<td>6</td>
</tr>
<tr>
<td>AT2G27389</td>
<td>2</td>
<td>AT2G27390</td>
<td>0</td>
</tr>
<tr>
<td>AT2G28670</td>
<td>7</td>
<td>AT2G28671</td>
<td>0</td>
</tr>
<tr>
<td>AT2G30766</td>
<td>2</td>
<td>AT2G30760</td>
<td>1</td>
</tr>
<tr>
<td>AT2G36026</td>
<td>8</td>
<td>AT2G36030</td>
<td>4</td>
</tr>
<tr>
<td>AT2G39518</td>
<td>8</td>
<td>AT2G39520</td>
<td>0</td>
</tr>
<tr>
<td>AT2G41230</td>
<td>2</td>
<td>AT2G41231</td>
<td>2</td>
</tr>
<tr>
<td>AT3G06868</td>
<td>6</td>
<td>AT3G06870</td>
<td>0</td>
</tr>
<tr>
<td>AT3G13404</td>
<td>0</td>
<td>AT3G13403</td>
<td>0</td>
</tr>
<tr>
<td>AT3G17670</td>
<td>9</td>
<td>AT3G17675</td>
<td>7</td>
</tr>
<tr>
<td>AT3G20898</td>
<td>7</td>
<td>AT3G20900</td>
<td>1</td>
</tr>
<tr>
<td>AT3G22190</td>
<td>8</td>
<td>AT3G22183</td>
<td>0</td>
</tr>
<tr>
<td>AT3G25013</td>
<td>9</td>
<td>AT3G25014</td>
<td>0</td>
</tr>
<tr>
<td>AT3G25560</td>
<td>8</td>
<td>AT3G25550</td>
<td>8</td>
</tr>
<tr>
<td>AT3G27415</td>
<td>2</td>
<td>AT3G27416</td>
<td>2</td>
</tr>
<tr>
<td>AT3G27810</td>
<td>6</td>
<td>AT3G27809</td>
<td>0</td>
</tr>
<tr>
<td>AT3G48300</td>
<td>6</td>
<td>AT3G48298</td>
<td>0</td>
</tr>
<tr>
<td>AT3G52310</td>
<td>9</td>
<td>AT3G52302</td>
<td>0</td>
</tr>
<tr>
<td>AT4G02200</td>
<td>8</td>
<td>AT4G02210</td>
<td>6</td>
</tr>
<tr>
<td>AT4G12382</td>
<td>6</td>
<td>AT4G12380</td>
<td>0</td>
</tr>
<tr>
<td>AT4G16410</td>
<td>8</td>
<td>AT4G16400</td>
<td>5</td>
</tr>
<tr>
<td>AT4G18501</td>
<td>4</td>
<td>AT4G18500</td>
<td>0</td>
</tr>
<tr>
<td>AT4G18720</td>
<td>9</td>
<td>AT4G18710</td>
<td>8</td>
</tr>
<tr>
<td>AT4G19100</td>
<td>9</td>
<td>AT4G19095</td>
<td>0</td>
</tr>
<tr>
<td>AT4G21930</td>
<td>8</td>
<td>AT4G21926</td>
<td>0</td>
</tr>
<tr>
<td>AT4G22513</td>
<td>6</td>
<td>AT4G22510</td>
<td>0</td>
</tr>
<tr>
<td>AT4G22517</td>
<td>6</td>
<td>AT4G22510</td>
<td>0</td>
</tr>
<tr>
<td>AT4G24026</td>
<td>7</td>
<td>AT4G24030</td>
<td>2</td>
</tr>
<tr>
<td>AT4G26288</td>
<td>8</td>
<td>AT4G26290</td>
<td>0</td>
</tr>
<tr>
<td>AT4G35440</td>
<td>9</td>
<td>AT4G35430</td>
<td>0</td>
</tr>
<tr>
<td>AT4G37410</td>
<td>7</td>
<td>AT4G37409</td>
<td>4</td>
</tr>
<tr>
<td>AT4G37480</td>
<td>9</td>
<td>AT4G37483</td>
<td>0</td>
</tr>
<tr>
<td>AT4G38825</td>
<td>5</td>
<td>AT4G38820</td>
<td>4</td>
</tr>
<tr>
<td>AT5G01580</td>
<td>9</td>
<td>AT5G01570</td>
<td>6</td>
</tr>
<tr>
<td>AT5G07571</td>
<td>7</td>
<td>AT5G07572</td>
<td>2</td>
</tr>
<tr>
<td>AT5G15580</td>
<td>8</td>
<td>AT5G15581</td>
<td>2</td>
</tr>
<tr>
<td>AT5G18210</td>
<td>9</td>
<td>AT5G18220</td>
<td>8</td>
</tr>
<tr>
<td>AT5G27945</td>
<td>9</td>
<td>AT5G27944</td>
<td>8</td>
</tr>
<tr>
<td>AT5G28920</td>
<td>2</td>
<td>AT5G28919</td>
<td>2</td>
</tr>
<tr>
<td>AT5G37415</td>
<td>6</td>
<td>AT5G37420</td>
<td>2</td>
</tr>
<tr>
<td>AT5G48205</td>
<td>9</td>
<td>AT5G48200</td>
<td>0</td>
</tr>
<tr>
<td>AT5G51120</td>
<td>9</td>
<td>AT5G51110</td>
<td>9</td>
</tr>
<tr>
<td>AT5G61490</td>
<td>8</td>
<td>AT5G61495</td>
<td>0</td>
</tr>
<tr>
<td>AT5G65610</td>
<td>2</td>
<td>AT5G65609</td>
<td>1</td>
</tr>
<tr>
<td>ATCG00280</td>
<td>9</td>
<td>ATCG00270</td>
<td>7</td>
</tr>
<tr>
<td>ATCG00480</td>
<td>9</td>
<td>ATCG00470</td>
<td>9</td>
</tr>
<tr>
<td>ATCG00810</td>
<td>9</td>
<td>ATCG00800</td>
<td>8</td>
</tr>
<tr>
<td>ATCG10000</td>
<td>8</td>
<td>ATCG01010</td>
<td>7</td>
</tr>
<tr>
<td>ATMG00040</td>
<td>9</td>
<td>ATMG00050</td>
<td>0</td>
</tr>
<tr>
<td>ATMG00090</td>
<td>9</td>
<td>ATMG00080</td>
<td>9</td>
</tr>
<tr>
<td>ATMG00310</td>
<td>6</td>
<td>ATMG00320</td>
<td>6</td>
</tr>
<tr>
<td>ATMG00490</td>
<td>6</td>
<td>ATMG00500</td>
<td>2</td>
</tr>
<tr>
<td>ATMG00880</td>
<td>1</td>
<td>ATMG00870</td>
<td>0</td>
</tr>
<tr>
<td>ATMG00990</td>
<td>9</td>
<td>ATMG01000</td>
<td>0</td>
</tr>
<tr>
<td>ATMG01010</td>
<td>6</td>
<td>ATMG01000</td>
<td>0</td>
</tr>
<tr>
<td>ATMG01230</td>
<td>0</td>
<td>ATMG01240</td>
<td>0</td>
</tr>
</tbody>
</table>
Supplementary Figure 3.1: A general methodology for the production of commercial hybrids by using cytoplasmic male sterility system.
Supplementary Figure 3.2: Graph represents GC content distribution in mitochondrial genomes of ICPA 2039, ICPB 2039, ICPH 2433 and ICPW 29 in red, blue, yellow and black colors respectively.
Supplementary Figure 3.3: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Triticum aestivum*. 
Supplemental Figure 3.4: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Zea mays*.  

![Gene order correlation diagram between *Cajanus cajan* and *Zea mays*](image-url)
Supplementary Figure 3.5: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Oryza sativa.*
Supplementary Figure 3.6: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Arabidopsis thaliana*. 
Supplementary Figure 3.7: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Beta vulgaris*.
Supplementary Figure 3.8: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Citrullus lanatus*.
Supplemental Figure 3.9: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Cucurbita pepo*.
Supplementary Figure 3.10: Correlation of gene order between the mitochondrial gene maps of Cajanus cajan and Nicotina tabacum.
Supplementary Figure 3.11: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Vitis vinifera*. 
Supplemental Figure 3.12: Correlation of gene order between the mitochondrial gene maps of *Cajanus cajan* and *Cucumis sativus*.
Supplementary Figure 3.13: Output of GS Reference Mapper 2.5, which represents rearrangements and no coverage regions identified after aligning additional sequencing reads of ICPB 2039 and ICPW 29 to the final assembly of ICPA 2039.
Supplementary Table 3.1: Details on 24 primer pairs designed from the end-sequences of the scaffolds to confirm the orientation in the mitochondrial genome assembly of ICPA 2039.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Forward primer (5’-3’)</th>
<th>Reverse primer (5’-3’)</th>
<th>Tm (°C)</th>
<th>Product size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*CcMtM01</td>
<td>AGGAGACAGCTGA</td>
<td>CTTGGGGGATAC</td>
<td>60.017</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>TGGTGCT</td>
<td>ACCTTCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM02</td>
<td>CCCAGCTAAAGCGA</td>
<td>ACTGGTGCTTACA</td>
<td>60.481</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>ATCTCA</td>
<td>CGAAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM03</td>
<td>TATCTTGCTTGAGA</td>
<td>GAACTCAAAGGC</td>
<td>59.767</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>GGGGAAG</td>
<td>GAGCCAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM04</td>
<td>TAGCGAAGGAGAG</td>
<td>GCAATGATTTTCAG</td>
<td>60.088</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>GGTGAGA</td>
<td>ACGCGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM05</td>
<td>GGAGGTTGAGGTCT</td>
<td>TGAAGCTAGGCCT</td>
<td>59.534</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>GTGTC</td>
<td>TATTGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM06</td>
<td>AGGCCCAACAACCT</td>
<td>AAGGTACGAGGTC</td>
<td>60.345</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>AAAAGG</td>
<td>GCTCAAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM07</td>
<td>TCTGAGTTTCTTTGG</td>
<td>TTTCTCAGGGAT</td>
<td>59.844</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>TCGGTC</td>
<td>GAGGTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM08</td>
<td>TAAAGCACTGTGTCG</td>
<td>CGCCACACCTCATC</td>
<td>59.871</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>TCACTC</td>
<td>ATATCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM09</td>
<td>GAGCATAAGGATTA</td>
<td>CAAGTGCTACGAG</td>
<td>59.411</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>GCAGAATCG</td>
<td>TGCTTCTCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM10</td>
<td>TGGGTCAGGAAAGGAC</td>
<td>AGAAAAATCTGCCT</td>
<td>59.959</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>AGGTAG</td>
<td>CCAGCAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM11</td>
<td>CCAACCTCTCGAAA</td>
<td>TCAGATGATTTGT</td>
<td>60.044</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>TCCAAA</td>
<td>GGACGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM12</td>
<td>ACAGACCAAGCAA</td>
<td>AGGGGCTAAAGGG</td>
<td>59.875</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>GGGCCTTA</td>
<td>AGTTTCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM13</td>
<td>CTGAGAGAAGGCTC</td>
<td>ATTATTTCCACCC</td>
<td>59.986</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td>GTGACC</td>
<td>CTCGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM14</td>
<td>ACCCCACAATCAGGC</td>
<td>CGAGGTCTCAACG</td>
<td>59.973</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>AAGTTTC</td>
<td>AAAGGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM15</td>
<td>ACTCCCGTGCAAGGAA</td>
<td>GAGCGCGCCTGAA</td>
<td>59.455</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>AGACTC</td>
<td>TAAGTGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM16</td>
<td>CGGGGTTGTGATAGT</td>
<td>CAATCCCTCTTCTG</td>
<td>59.993</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>TGCAGT</td>
<td>GAAGGCG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM17</td>
<td>AAAAGCGTGCTCAGCC</td>
<td>TAAAGGGAAGGCT</td>
<td>59.853</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>TTCAAGA</td>
<td>CGACGAA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM18</td>
<td>TTCCACCTTGCTTTCT</td>
<td>AGTTGCTGAAACC</td>
<td>60.132</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>CGCTT</td>
<td>TGAGCGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM19</td>
<td>CTCCTTCCACGGGA</td>
<td>AAGCAAATAGATA</td>
<td>60.042</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>TGAACA</td>
<td>GCCCCCCCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM20</td>
<td>TGCTCTCTTGCTAC</td>
<td>ACGCCTGGTCCATC</td>
<td>59.978</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>GGGATT</td>
<td>ACTTACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM21</td>
<td>CTCGAGGGAGGAAAG</td>
<td>TATCTCCCCCTTGC</td>
<td>59.122</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>ATAAGA</td>
<td>CTTTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM22</td>
<td>CAGAACGAATCTCCT</td>
<td>ACTCAGTCTAC</td>
<td>58.941</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>TATCGCC</td>
<td>CCCGCTCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM23</td>
<td>GTTTCGGTACGAAA</td>
<td>TCCGTCTTGCCTTT</td>
<td>60.249</td>
<td>202</td>
</tr>
<tr>
<td>---------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>GCCTGA</td>
<td>GCTACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CcMtM24</td>
<td>AGTCCCCGGGATCGT</td>
<td>TGTCACAAGGAAG</td>
<td>59.827</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>AAAGTT</td>
<td>GTGGTCA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*CcMtM: *Cajanus cajan* mitochondria marker.
Supplementary Table 3.2: Genes in pigeonpea mitochondrial genome (ICPA2039).

<table>
<thead>
<tr>
<th>Genes</th>
<th>Size</th>
<th>Coordinates</th>
<th>Strand</th>
<th>Organism with top hit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>I. Complex I genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nad1a</td>
<td>383</td>
<td>82256</td>
<td>81873</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad1b</td>
<td>88</td>
<td>302673</td>
<td>302761</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad1c</td>
<td>194</td>
<td>303623</td>
<td>303817</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad1d</td>
<td>86</td>
<td>371688</td>
<td>371774</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad1e</td>
<td>254</td>
<td>374878</td>
<td>375132</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad1</td>
<td>1005</td>
<td>-</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>nad2a</td>
<td>152</td>
<td>175842</td>
<td>175994</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad2b</td>
<td>401</td>
<td>177235</td>
<td>177636</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad2c</td>
<td>158</td>
<td>335490</td>
<td>335332</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad2d</td>
<td>638</td>
<td>332814</td>
<td>332176</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad2e</td>
<td>170</td>
<td>330741</td>
<td>330571</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad2</td>
<td>1519</td>
<td>-</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>nad3</td>
<td>353</td>
<td>108269</td>
<td>108622</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad4a</td>
<td>458</td>
<td>351289</td>
<td>351747</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad4b</td>
<td>584</td>
<td>353136</td>
<td>353720</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad4c</td>
<td>419</td>
<td>356763</td>
<td>357182</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad4d</td>
<td>86</td>
<td>359741</td>
<td>359827</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad4</td>
<td>1547</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>nad4L</td>
<td>296</td>
<td>294002</td>
<td>294271</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad5a</td>
<td>236</td>
<td>45291</td>
<td>45055</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad5b</td>
<td>1232</td>
<td>47332</td>
<td>46100</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad5c</td>
<td>395</td>
<td>313079</td>
<td>313474</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad5d</td>
<td>137</td>
<td>314387</td>
<td>314524</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>nad5</td>
<td>2000</td>
<td>-</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>nad6</td>
<td>614</td>
<td>79231</td>
<td>78617</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad7a</td>
<td>161</td>
<td>164750</td>
<td>164589</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad7b</td>
<td>101</td>
<td>163739</td>
<td>163638</td>
<td>- Cucurbita pepo</td>
</tr>
<tr>
<td>nad7c</td>
<td>464</td>
<td>162302</td>
<td>161838</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad7d</td>
<td>245</td>
<td>160788</td>
<td>160543</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad7e</td>
<td>323</td>
<td>158734</td>
<td>158411</td>
<td>- Vigna radiata</td>
</tr>
<tr>
<td>nad7</td>
<td>1294</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>nad9</td>
<td>569</td>
<td>433769</td>
<td>434338</td>
<td>+ Vigna radiata</td>
</tr>
<tr>
<td>II. Complex II genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdh4</td>
<td>239</td>
<td>261724</td>
<td>261963</td>
<td>+ Citrullus lanatus</td>
</tr>
<tr>
<td>III. Complex III genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cob</td>
<td>1172</td>
<td>319048</td>
<td>320220</td>
<td>+ Cycas taitungensis</td>
</tr>
<tr>
<td>IV. Complex IV genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td>1579</td>
<td>299624</td>
<td>301203</td>
<td>+ Vigna radiata</td>
</tr>
</tbody>
</table>
V. Complex V genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>atp1</td>
<td>1511</td>
<td>233337</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>atp4</td>
<td>566</td>
<td>294470</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>atp6</td>
<td>728</td>
<td>301</td>
<td>+</td>
<td>Carica papaya</td>
</tr>
<tr>
<td>atp8</td>
<td>482</td>
<td>243540</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>atp9</td>
<td>170</td>
<td>468828</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
</tbody>
</table>

VI. Cytochrome c biogenesis genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>ccmB</td>
<td>620</td>
<td>83549</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>ccmC</td>
<td>740</td>
<td>34816</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>ccmFCa</td>
<td>779</td>
<td>15633</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>ccmFCb</td>
<td>566</td>
<td>11352</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>ccmFC</td>
<td>1345</td>
<td></td>
<td>x</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>ccmFn</td>
<td>1733</td>
<td>37237</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
</tbody>
</table>

VII. Ribosomal protein genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpl16</td>
<td>510</td>
<td>308657</td>
<td>+</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>rpl5</td>
<td>554</td>
<td>317272</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps1</td>
<td>620</td>
<td>44252</td>
<td>+</td>
<td>Carica papaya</td>
</tr>
<tr>
<td>rps10a</td>
<td>309</td>
<td>296174</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps10b</td>
<td>107</td>
<td>299315</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps10</td>
<td>416</td>
<td></td>
<td>x</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps12</td>
<td>374</td>
<td>108674</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps14</td>
<td>299</td>
<td>317837</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps19a</td>
<td>68</td>
<td>507787</td>
<td>-</td>
<td>Cucurbita pepo</td>
</tr>
<tr>
<td>rps19b</td>
<td>119</td>
<td>305180</td>
<td>+</td>
<td>Cucurbita pepo</td>
</tr>
<tr>
<td>rps19</td>
<td>187</td>
<td></td>
<td>x</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps3a</td>
<td>74</td>
<td>305316</td>
<td>+</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td>rps3b</td>
<td>1571</td>
<td>307156</td>
<td>+</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td>rps3</td>
<td>1645</td>
<td></td>
<td>x</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps4</td>
<td>1037</td>
<td>60663</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps7</td>
<td>89</td>
<td>34089</td>
<td>-</td>
<td>Citrullus lanatus</td>
</tr>
</tbody>
</table>

VIII. Other protein coding genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>matR</td>
<td>2017</td>
<td>372328</td>
<td>+</td>
<td>Carica papaya</td>
</tr>
<tr>
<td>mttB</td>
<td>722</td>
<td>395960</td>
<td>-</td>
<td>Nicotiana tabacum</td>
</tr>
</tbody>
</table>

IX. rRNA genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>rrn5</td>
<td>113</td>
<td>441424</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rrnL</td>
<td>3154</td>
<td>122673</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rrnS</td>
<td>1988</td>
<td>443580</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
</tbody>
</table>

X. tRNA genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>trnN</td>
<td>71</td>
<td>504005</td>
<td>-</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td>trnD</td>
<td>73</td>
<td>349884</td>
<td>-</td>
<td>Nicotiana tabacum</td>
</tr>
<tr>
<td>trnC-1</td>
<td>72</td>
<td>250520</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Feature</td>
<td>Start</td>
<td>End</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>-------</td>
<td>--------------------------</td>
<td></td>
</tr>
<tr>
<td>trnC-2</td>
<td>70</td>
<td>383649 383579</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnQ</td>
<td>71</td>
<td>381197 381126</td>
<td>Cucumis sativus</td>
<td></td>
</tr>
<tr>
<td>trnE</td>
<td>71</td>
<td>377238 377309</td>
<td>Arabidopsis thaliana</td>
<td></td>
</tr>
<tr>
<td>trnG</td>
<td>71</td>
<td>507338 507267</td>
<td>Beta vulgaris</td>
<td></td>
</tr>
<tr>
<td>trnH</td>
<td>73</td>
<td>252351 252278</td>
<td>Cucumis sativus</td>
<td></td>
</tr>
<tr>
<td>trnK-1</td>
<td>72</td>
<td>421642 421570</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnK-2</td>
<td>72</td>
<td>526879 526807</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnM-1</td>
<td>73</td>
<td>119146 119073</td>
<td>Nicotiana tabacum</td>
<td></td>
</tr>
<tr>
<td>trnM-2</td>
<td>73</td>
<td>190371 190444</td>
<td>Nicotiana tabacum</td>
<td></td>
</tr>
<tr>
<td>trnM-3</td>
<td>73</td>
<td>243291 243364</td>
<td>Nicotiana tabacum</td>
<td></td>
</tr>
<tr>
<td>trnM-4</td>
<td>81</td>
<td>368552 368471</td>
<td>Zea mays</td>
<td></td>
</tr>
<tr>
<td>trnM-5-cp</td>
<td>72</td>
<td>378147 378219</td>
<td>Citrullus lanatus</td>
<td></td>
</tr>
<tr>
<td>trnF</td>
<td>73</td>
<td>194491 194564</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnP</td>
<td>74</td>
<td>194824 194898</td>
<td>Cucumis sativus</td>
<td></td>
</tr>
<tr>
<td>trnS</td>
<td>87</td>
<td>194042 194129</td>
<td>Zea perennis</td>
<td></td>
</tr>
<tr>
<td>trnW-cp</td>
<td>73</td>
<td>445511 445438</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnY</td>
<td>82</td>
<td>503562 503480</td>
<td>Vigna radiata</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3.3: Gene annotation of scaffolds derived from mitochondrial sequences of ICPH 2433.

<table>
<thead>
<tr>
<th>ScaffoldID</th>
<th>Genes</th>
<th>Size</th>
<th>Coordinates</th>
<th>Strand</th>
<th>Organism with top hit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>Scaffold1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Scaffold101</td>
<td>Beta vulgaris</td>
<td>Beta vulgaris</td>
<td>Citrullus lanatus</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>---------------</td>
<td>---------------</td>
<td>------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>atp4</td>
<td>569</td>
<td>32026</td>
<td>32595</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>atp9</td>
<td>65</td>
<td>13907</td>
<td>13842</td>
<td>-</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td>cob</td>
<td>1172</td>
<td>56638</td>
<td>57810</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>coxl</td>
<td>1580</td>
<td>37187</td>
<td>38767</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>matR</td>
<td>2002</td>
<td>109758</td>
<td>111760</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>mtB</td>
<td>722</td>
<td>133343</td>
<td>132621</td>
<td>-</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td>nad1b</td>
<td>92</td>
<td>40244</td>
<td>40336</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad1c</td>
<td>194</td>
<td>41196</td>
<td>41390</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad1e</td>
<td>254</td>
<td>112252</td>
<td>112506</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad2c</td>
<td>158</td>
<td>73097</td>
<td>72939</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad2d</td>
<td>638</td>
<td>70419</td>
<td>69781</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad2e</td>
<td>185</td>
<td>68357</td>
<td>68172</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad4a</td>
<td>458</td>
<td>88644</td>
<td>89102</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad4b</td>
<td>584</td>
<td>90493</td>
<td>91077</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad4c</td>
<td>419</td>
<td>94122</td>
<td>94541</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad4d</td>
<td>86</td>
<td>97105</td>
<td>97191</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad5c</td>
<td>395</td>
<td>50665</td>
<td>51060</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>nad5d</td>
<td>137</td>
<td>51973</td>
<td>52110</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rpl16</td>
<td>535</td>
<td>46215</td>
<td>46750</td>
<td>+</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>rpl5</td>
<td>554</td>
<td>54859</td>
<td>55413</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps10a</td>
<td>248</td>
<td>33793</td>
<td>34041</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps10b</td>
<td>119</td>
<td>36865</td>
<td>36984</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps14</td>
<td>299</td>
<td>55424</td>
<td>55723</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps19b</td>
<td>119</td>
<td>42756</td>
<td>42875</td>
<td>+</td>
<td>Cucurbita pepo</td>
</tr>
<tr>
<td>rps3a</td>
<td>74</td>
<td>42892</td>
<td>42966</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>rps3b</td>
<td>1583</td>
<td>44729</td>
<td>46312</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>sdh4</td>
<td>62</td>
<td>64990</td>
<td>64928</td>
<td>-</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td>trnM</td>
<td>72</td>
<td>115524</td>
<td>115596</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>trnC</td>
<td>70</td>
<td>120961</td>
<td>121031</td>
<td>-</td>
<td>Cucumis sativus,</td>
</tr>
<tr>
<td>trnQ</td>
<td>71</td>
<td>118507</td>
<td>118578</td>
<td>-</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>trnE</td>
<td>71</td>
<td>114612</td>
<td>114683</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>trnK</td>
<td>72</td>
<td>159140</td>
<td>159212</td>
<td>-</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td>trnY</td>
<td>82</td>
<td>32021</td>
<td>32103</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
</tbody>
</table>
## Supplementary Table 3.4: Gene annotation of scaffolds derived from mitochondrial sequences of ICPW 29.

<table>
<thead>
<tr>
<th>ScaffoldID</th>
<th>Genes</th>
<th>Size</th>
<th>Coordinates From</th>
<th>Coordinates To</th>
<th>Strand</th>
<th>Organism with top hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffold1</td>
<td>atp1</td>
<td>1511</td>
<td>98948</td>
<td>97437</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>ccmB</td>
<td>617</td>
<td>77628</td>
<td>78245</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>cob</td>
<td>1172</td>
<td>148380</td>
<td>149552</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>cox3</td>
<td>794</td>
<td>108456</td>
<td>109250</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>matR</td>
<td>2003</td>
<td>202103</td>
<td>204106</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>mttB</td>
<td>722</td>
<td>225652</td>
<td>224930</td>
<td>-</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td></td>
<td>nad1a</td>
<td>383</td>
<td>78921</td>
<td>79304</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad1d</td>
<td>95</td>
<td>201394</td>
<td>201489</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad1e</td>
<td>254</td>
<td>204598</td>
<td>204852</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad2c</td>
<td>158</td>
<td>164840</td>
<td>164682</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad2d</td>
<td>638</td>
<td>162161</td>
<td>161523</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad2e</td>
<td>185</td>
<td>160099</td>
<td>159914</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad3</td>
<td>353</td>
<td>52916</td>
<td>52563</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad4a</td>
<td>458</td>
<td>180979</td>
<td>181437</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad4b</td>
<td>584</td>
<td>182829</td>
<td>183413</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad4c</td>
<td>419</td>
<td>186460</td>
<td>186879</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad4d</td>
<td>86</td>
<td>189445</td>
<td>189531</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad5c</td>
<td>395</td>
<td>142404</td>
<td>142799</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad5d</td>
<td>137</td>
<td>143712</td>
<td>143849</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad6</td>
<td>614</td>
<td>81947</td>
<td>82561</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rpl5</td>
<td>554</td>
<td>146598</td>
<td>147152</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rps12</td>
<td>374</td>
<td>52511</td>
<td>52137</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rps14</td>
<td>299</td>
<td>147163</td>
<td>147462</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>sdh4</td>
<td>239</td>
<td>109181</td>
<td>109420</td>
<td>+</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td></td>
<td>trnD</td>
<td>73</td>
<td>179499</td>
<td>179572</td>
<td>-</td>
<td>Nicotiana tabacum</td>
</tr>
<tr>
<td></td>
<td>trnC</td>
<td>70</td>
<td>213262</td>
<td>213332</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>trnQ</td>
<td>71</td>
<td>210808</td>
<td>210879</td>
<td>-</td>
<td>Cucumis sativus Arabidopsis thaliana</td>
</tr>
<tr>
<td></td>
<td>trnE</td>
<td>71</td>
<td>206958</td>
<td>207029</td>
<td>+</td>
<td>Nicotiana tabacum</td>
</tr>
<tr>
<td></td>
<td>trnM</td>
<td>72</td>
<td>41732</td>
<td>41804</td>
<td>+</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td></td>
<td>trnM</td>
<td>72</td>
<td>207870</td>
<td>207942</td>
<td>+</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>trnF</td>
<td>73</td>
<td>241150</td>
<td>241223</td>
<td>-</td>
<td>Cucumis sativus</td>
</tr>
<tr>
<td></td>
<td>trnP</td>
<td>74</td>
<td>240817</td>
<td>240891</td>
<td>-</td>
<td>Zea perennis</td>
</tr>
<tr>
<td></td>
<td>trnS</td>
<td>87</td>
<td>241585</td>
<td>241672</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold2</td>
<td>ccmC</td>
<td>737</td>
<td>10172</td>
<td>9435</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>ccmFn</td>
<td>1733</td>
<td>12593</td>
<td>10860</td>
<td>-</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Gene</td>
<td>Scaffold</td>
<td>Accession</td>
<td>3' End</td>
<td>5' Start</td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>nad5b</td>
<td>Scaffold3</td>
<td>1193</td>
<td>21696</td>
<td>22889</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>nad7a</td>
<td>Scaffold3</td>
<td>161</td>
<td>142571</td>
<td>142410</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>nad7b</td>
<td>Scaffold3</td>
<td>101</td>
<td>141560</td>
<td>141459</td>
<td>Cucurbita pepo</td>
<td></td>
</tr>
<tr>
<td>nad7c</td>
<td>Scaffold3</td>
<td>464</td>
<td>140122</td>
<td>139658</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>nad7d</td>
<td>Scaffold3</td>
<td>245</td>
<td>138608</td>
<td>138363</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>nad7e</td>
<td>Scaffold3</td>
<td>323</td>
<td>136551</td>
<td>136228</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>nad9</td>
<td>Scaffold3</td>
<td>491</td>
<td>93404</td>
<td>93895</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rps1</td>
<td>Scaffold3</td>
<td>620</td>
<td>19608</td>
<td>20228</td>
<td>Carica papaya</td>
<td></td>
</tr>
<tr>
<td>rps12</td>
<td>Scaffold3</td>
<td>149</td>
<td>42211</td>
<td>42360</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rps14</td>
<td>Scaffold3</td>
<td>68</td>
<td>36215</td>
<td>36147</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rps19a</td>
<td>Scaffold3</td>
<td>77</td>
<td>64757</td>
<td>64680</td>
<td>Carica papaya</td>
<td></td>
</tr>
<tr>
<td>rps19b</td>
<td>Scaffold3</td>
<td>59</td>
<td>156654</td>
<td>156713</td>
<td>Carica papaya</td>
<td></td>
</tr>
<tr>
<td>rps4</td>
<td>Scaffold3</td>
<td>1037</td>
<td>35953</td>
<td>34916</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rps7</td>
<td>Scaffold3</td>
<td>74</td>
<td>9445</td>
<td>9371</td>
<td>Citrullus lanatus</td>
<td></td>
</tr>
<tr>
<td>trnN</td>
<td>Scaffold3</td>
<td>71</td>
<td>60615</td>
<td>60686</td>
<td>Beta vulgaris</td>
<td></td>
</tr>
<tr>
<td>trnC</td>
<td>Scaffold3</td>
<td>72</td>
<td>151173</td>
<td>151245</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnG</td>
<td>Scaffold3</td>
<td>72</td>
<td>64237</td>
<td>64309</td>
<td>Beta vulgaris</td>
<td></td>
</tr>
<tr>
<td>trnH</td>
<td>Scaffold3</td>
<td>75</td>
<td>149413</td>
<td>149488</td>
<td>Beta vulgaris</td>
<td></td>
</tr>
<tr>
<td>trnK</td>
<td>Scaffold3</td>
<td>72</td>
<td>84781</td>
<td>84853</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnW-cp</td>
<td>Scaffold3</td>
<td>73</td>
<td>104998</td>
<td>105071</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>trnY</td>
<td>Scaffold3</td>
<td>82</td>
<td>60161</td>
<td>60243</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>Scaffold9</td>
<td></td>
<td>194</td>
<td>847</td>
<td>1041</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rpl16</td>
<td>Scaffold9</td>
<td>511</td>
<td>5890</td>
<td>6401</td>
<td>Arabidopsis thaliana</td>
<td></td>
</tr>
<tr>
<td>rps19b</td>
<td>Scaffold9</td>
<td>119</td>
<td>2407</td>
<td>2526</td>
<td>Cucurbita pepo</td>
<td></td>
</tr>
<tr>
<td>rps3a</td>
<td>Scaffold9</td>
<td>74</td>
<td>2543</td>
<td>2617</td>
<td>Vigna radiata</td>
<td></td>
</tr>
<tr>
<td>rps3b</td>
<td>Scaffold9</td>
<td>1583</td>
<td>4380</td>
<td>5963</td>
<td>Vigna radiata</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3.5: Gene annotation of scaffolds derived from mitochondrial sequences of ICPB 2039.

<table>
<thead>
<tr>
<th>ScaffoldID</th>
<th>Genes</th>
<th>Size</th>
<th>Coordinates</th>
<th>Strand</th>
<th>Organism with top hit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>Scaffold10</td>
<td>nad2a</td>
<td>152</td>
<td>2143</td>
<td>1991</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad2b</td>
<td>401</td>
<td>750</td>
<td>349</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rps10</td>
<td>47</td>
<td>2296</td>
<td>2249</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold13</td>
<td>nad3</td>
<td>77</td>
<td>180</td>
<td>103</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rps12</td>
<td>173</td>
<td>6067</td>
<td>6240</td>
<td>Brassica napus</td>
</tr>
<tr>
<td></td>
<td>rps14</td>
<td>68</td>
<td>96</td>
<td>28</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold1</td>
<td>nad3</td>
<td>353</td>
<td>14036</td>
<td>13683</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>rps12</td>
<td>374</td>
<td>13631</td>
<td>13257</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold30</td>
<td>rps19</td>
<td>77</td>
<td>4068</td>
<td>3991</td>
<td>Carica papaya</td>
</tr>
<tr>
<td></td>
<td>trnN</td>
<td>71</td>
<td>216</td>
<td>287</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td></td>
<td>trnG</td>
<td>71</td>
<td>3548</td>
<td>3619</td>
<td>Beta vulgaris</td>
</tr>
<tr>
<td>Scaffold31</td>
<td>ccmFn</td>
<td>455</td>
<td>458</td>
<td>3</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold33</td>
<td>atp9</td>
<td>221</td>
<td>3980</td>
<td>3759</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold34</td>
<td>matR</td>
<td>164</td>
<td>1616</td>
<td>1452</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>nad1</td>
<td>254</td>
<td>960</td>
<td>706</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>trnE</td>
<td>70</td>
<td>4851</td>
<td>4921</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td></td>
<td>trnM</td>
<td>72</td>
<td>3938</td>
<td>4010</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td>Scaffold3</td>
<td>ccmB</td>
<td>59</td>
<td>1956</td>
<td>2015</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>cox3</td>
<td>794</td>
<td>7881</td>
<td>7087</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>sdh4</td>
<td>323</td>
<td>7156</td>
<td>6833</td>
<td>Citrullus lanatus</td>
</tr>
<tr>
<td>Scaffold45</td>
<td>nad4</td>
<td>419</td>
<td>2112</td>
<td>1693</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold49</td>
<td>nad2c</td>
<td>158</td>
<td>1280</td>
<td>1438</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold7</td>
<td>nad9</td>
<td>570</td>
<td>15190</td>
<td>14620</td>
<td>Nicotiana tabacum</td>
</tr>
<tr>
<td>Scaffold16</td>
<td>trnK</td>
<td>72</td>
<td>1549</td>
<td>1621</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold17</td>
<td>trnW-cp</td>
<td>73</td>
<td>7159</td>
<td>7232</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold25</td>
<td>trnM</td>
<td>71</td>
<td>3331</td>
<td>3402</td>
<td>Zea mays</td>
</tr>
<tr>
<td>Scaffold36</td>
<td>trnW-cp</td>
<td>73</td>
<td>7159</td>
<td>7232</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td>Scaffold5</td>
<td>trnF</td>
<td>73</td>
<td>17592</td>
<td>17665</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>trnP</td>
<td>74</td>
<td>17260</td>
<td>17334</td>
<td>Cucumis sativus</td>
</tr>
<tr>
<td></td>
<td>trnS</td>
<td>87</td>
<td>18027</td>
<td>18114</td>
<td>Zea perennis</td>
</tr>
<tr>
<td>Scaffold6</td>
<td>trnC</td>
<td>70</td>
<td>10058</td>
<td>10128</td>
<td>Vigna radiata</td>
</tr>
<tr>
<td></td>
<td>trnQ</td>
<td>71</td>
<td>13392</td>
<td>13463</td>
<td>Cucumis sativus</td>
</tr>
</tbody>
</table>
Supplementary Table 3.6: Genomic rearrangements observed in ICPA 2039 while comparing with mitochondrial sequences of ICPB 2039, ICPW 29 and ICPH 2433 separately.

<table>
<thead>
<tr>
<th>Rearrangement Regions From To</th>
<th>A/B</th>
<th>A/W</th>
<th>A/H</th>
</tr>
</thead>
<tbody>
<tr>
<td>3441 7072</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15709 18693</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34909 165239</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72614 403154</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>107485 219108</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>143366 144892</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>164987 165000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>170658 225317</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>187564 399514</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>191084 489849</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>193458 195972</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>249675 263641</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>251365 253166</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>252598 252974</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>262266 401325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>262352 263295</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>262339 401289</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>263124 470227</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>275554 471358</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>276013 470845</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>298472 298979</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>306210 307224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>342909 344446</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>379049 382176</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>433127 484083</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>443386 444100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>446611 451191</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3.7: No coverage regions observed in ICPA 2039 while comparing with mitochondrial sequences of ICPB 2039, ICPW 29 and ICPH 2433 separately.

<table>
<thead>
<tr>
<th>No Coverage Regions</th>
<th>A/B</th>
<th>A/W</th>
<th>A/H</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td>Length</td>
<td></td>
</tr>
<tr>
<td>489850</td>
<td>499083</td>
<td>9234</td>
<td>✓</td>
</tr>
<tr>
<td>476053</td>
<td>484470</td>
<td>8418</td>
<td>✓</td>
</tr>
<tr>
<td>212270</td>
<td>219837</td>
<td>7568</td>
<td>✓</td>
</tr>
<tr>
<td>165134</td>
<td>170282</td>
<td>5149</td>
<td>✓</td>
</tr>
<tr>
<td>419186</td>
<td>424114</td>
<td>4929</td>
<td>✓</td>
</tr>
<tr>
<td>257348</td>
<td>262265</td>
<td>4918</td>
<td>✓</td>
</tr>
<tr>
<td>531943</td>
<td>536388</td>
<td>4446</td>
<td>✓</td>
</tr>
<tr>
<td>265782</td>
<td>269725</td>
<td>3944</td>
<td>✓</td>
</tr>
<tr>
<td>69876</td>
<td>72608</td>
<td>2733</td>
<td>✓</td>
</tr>
<tr>
<td>411099</td>
<td>413552</td>
<td>2454</td>
<td>✓</td>
</tr>
<tr>
<td>309903</td>
<td>312076</td>
<td>2174</td>
<td>✓</td>
</tr>
<tr>
<td>430061</td>
<td>432077</td>
<td>2017</td>
<td>✓</td>
</tr>
<tr>
<td>157104</td>
<td>158668</td>
<td>1565</td>
<td>✓</td>
</tr>
<tr>
<td>81894</td>
<td>83428</td>
<td>1535</td>
<td>✓</td>
</tr>
<tr>
<td>324064</td>
<td>325537</td>
<td>1474</td>
<td>✓</td>
</tr>
<tr>
<td>263655</td>
<td>265027</td>
<td>1373</td>
<td>✓</td>
</tr>
<tr>
<td>222478</td>
<td>223535</td>
<td>1058</td>
<td>✓</td>
</tr>
<tr>
<td>489860</td>
<td>499011</td>
<td>9152</td>
<td>-</td>
</tr>
<tr>
<td>476087</td>
<td>484001</td>
<td>7915</td>
<td>-</td>
</tr>
<tr>
<td>212269</td>
<td>219101</td>
<td>6833</td>
<td>-</td>
</tr>
<tr>
<td>257348</td>
<td>262265</td>
<td>4918</td>
<td>-</td>
</tr>
<tr>
<td>419548</td>
<td>424117</td>
<td>4570</td>
<td>-</td>
</tr>
<tr>
<td>166121</td>
<td>170233</td>
<td>4113</td>
<td>-</td>
</tr>
<tr>
<td>265782</td>
<td>269725</td>
<td>3944</td>
<td>-</td>
</tr>
<tr>
<td>69913</td>
<td>72520</td>
<td>2608</td>
<td>-</td>
</tr>
<tr>
<td>411099</td>
<td>413510</td>
<td>2412</td>
<td>-</td>
</tr>
<tr>
<td>430061</td>
<td>432077</td>
<td>2017</td>
<td>-</td>
</tr>
<tr>
<td>310202</td>
<td>312008</td>
<td>1807</td>
<td>-</td>
</tr>
<tr>
<td>263655</td>
<td>265107</td>
<td>1453</td>
<td>-</td>
</tr>
<tr>
<td>217615</td>
<td>218657</td>
<td>1043</td>
<td>-</td>
</tr>
<tr>
<td>257653</td>
<td>261955</td>
<td>4303</td>
<td>-</td>
</tr>
</tbody>
</table>
Supplementary Table 5.1: List of validated set of 148 uniparentally-expressed genes.

<table>
<thead>
<tr>
<th>GeneId</th>
<th>Imprinting status</th>
<th>Age</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G13495</td>
<td>iPEG</td>
<td>0</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G35183</td>
<td>iPEG</td>
<td>0</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G77960</td>
<td>iMEG</td>
<td>1</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT3G49770</td>
<td>iPEG</td>
<td>1</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G51000</td>
<td>iMEG</td>
<td>1</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G61090</td>
<td>iMEG</td>
<td>1</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G23110</td>
<td>iPEG</td>
<td>2</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT3G11160</td>
<td>iPEG</td>
<td>2</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G42235</td>
<td>iMEG</td>
<td>2</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G54350</td>
<td>iPEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G07690</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G52460</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G12180</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT3G57250</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G46300</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G63960</td>
<td>iMEG</td>
<td>2</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G62110</td>
<td>iPEG</td>
<td>2</td>
<td>Gehring et al., 2009</td>
</tr>
<tr>
<td>AT1G20690</td>
<td>iMEG</td>
<td>3</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G11940</td>
<td>iPEG</td>
<td>5</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G22400</td>
<td>iMEG</td>
<td>5</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G20680</td>
<td>iMEG</td>
<td>5</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G11460</td>
<td>iPEG</td>
<td>6</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G33780</td>
<td>iPEG</td>
<td>6</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G60410</td>
<td>iMEG</td>
<td>6</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT3G11310</td>
<td>iMEG</td>
<td>6</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G59930</td>
<td>iMEG</td>
<td>6</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G53250</td>
<td>iMEG</td>
<td>6</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT2G17690</td>
<td>iMEG</td>
<td>6</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G35490</td>
<td>iMEG</td>
<td>6</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G49290</td>
<td>iPEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT3G62230</td>
<td>iPEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G60400</td>
<td>iPEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G11810</td>
<td>iPEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G61330</td>
<td>iPEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G19160</td>
<td>iMEG</td>
<td>6</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT2G36560</td>
<td>iPEG</td>
<td>7</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT4G05470</td>
<td>iPEG</td>
<td>7</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G42670</td>
<td>iPEG</td>
<td>7</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G17990</td>
<td>iMEG</td>
<td>7</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G51300</td>
<td>iMEG</td>
<td>7</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Type</td>
<td>Version</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>---------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>AT1G55050</td>
<td>iMEG</td>
<td>7</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G59670</td>
<td>iMEG</td>
<td>7</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT2G18880</td>
<td>iMEG</td>
<td>7</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G37290</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G66630</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G18250</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G28300</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT3G11490</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G70560</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G32370</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011; Gehring et al., 2009</td>
</tr>
<tr>
<td>AT1G67830</td>
<td>iPEG</td>
<td>8</td>
<td>Gehring et al., 2011; Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G18650</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G18150</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G53870</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G31060</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G00220</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G69900</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT2G28380</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G21790</td>
<td>iMEG</td>
<td>8</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G05440</td>
<td>iPEG</td>
<td>8</td>
<td>unpublished</td>
</tr>
<tr>
<td>AT1G34650</td>
<td>iPEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G11400</td>
<td>iPEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G23320</td>
<td>iPEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G01530</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT3G23060</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G73560</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G62340</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G20730</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G25530</td>
<td>iMEG</td>
<td>8</td>
<td>Wolff et al., 2011; Kinoshita et al., 2004</td>
</tr>
<tr>
<td>AT3G03260</td>
<td>iMEG</td>
<td>8</td>
<td>Gehring et al., 2009</td>
</tr>
<tr>
<td>AT5G17320</td>
<td>iMEG</td>
<td>8</td>
<td>Gehring et al., 2009</td>
</tr>
<tr>
<td>AT3G08040</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT4G13460</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G43780</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT3G16320</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G40520</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT4G15900</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G49600</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G20910</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G67820</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G57820</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G63020</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT5G53150</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G07705</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G57800</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G21450</td>
<td>iPEG</td>
<td>9</td>
<td>Gehring et al., 2011</td>
</tr>
</tbody>
</table>
AT3G14205 iPEG 9 Gehring et al., 2011
AT4G17486 iPEG 9 Gehring et al., 2011
AT1G47490 iPEG 9 Gehring et al., 2011
AT5G60760 iPEG 9 Gehring et al., 2011
AT1G60220 iPEG 9 Gehring et al., 2011
AT4G08590 iPEG 9 Gehring et al., 2011
AT1G59890 iPEG 9 Gehring et al., 2011
AT3G45090 iPEG 9 Gehring et al., 2011
AT5G60760 iPEG 9 Gehring et al., 2011
AT1G60220 iPEG 9 Gehring et al., 2011
AT4G08590 iPEG 9 Gehring et al., 2011
AT1G59890 iPEG 9 Gehring et al., 2011
AT3G45090 iPEG 9 Gehring et al., 2011
AT5G60760 iPEG 9 Gehring et al., 2011
AT1G17770 iPEG 9 Gehring et al., 2011; Wolff et al., 2011
AT1G48910 iPEG 9 Gehring et al., 2011; Wolff et al., 2011
AT2G31360 iMEG 9 Hsieh et al., 2011
AT5G03370 iMEG 9 Hsieh et al., 2011
AT1G76250 iMEG 9 Hsieh et al., 2011
AT3G22180 iMEG 9 Hsieh et al., 2011
AT3G17250 iMEG 9 Hsieh et al., 2011
AT5G47770 iMEG 9 Hsieh et al., 2011
AT4G39955 iMEG 9 Hsieh et al., 2011
AT1G77000 iMEG 9 Hsieh et al., 2011
AT2G34880 iMEG 9 Hsieh et al., 2011
AT1G62660 iMEG 9 Hsieh et al., 2011
AT5G22920 iMEG 9 Hsieh et al., 2011
AT1G76820 iMEG 9 Hsieh et al., 2011
AT5G47770 iMEG 9 Hsieh et al., 2011
AT4G16760 iMEG 9 Hsieh et al., 2011
AT4G01840 iMEG 9 Hsieh et al., 2011
AT1G08050 iMEG 9 Hsieh et al., 2011
AT5G02630 iMEG 9 Hsieh et al., 2011
AT5G03280 iMEG 9 Hsieh et al., 2011
AT2G35670 iMEG 9 Jullien et al., 2006
AT5G54650 iMEG 9 Gerald et al., 2009
AT3G19350 iMEG 9 Tiwari et al., 2008
AT1G02580 iMEG 9 Vielle-Calzada et al., 1999
AT1G60740 iPEG 9 Wolff et al., 2011
AT1G65360 iPEG 9 Wolff et al., 2011
AT4G10160 iPEG 9 Wolff et al., 2011
AT4G20800 iPEG 9 Wolff et al., 2011
AT3G19160 iPEG 9 Wolff et al., 2011
AT3G50720 iPEG 9 Wolff et al., 2011
AT5G15140 iPEG 9 Wolff et al., 2011
AT4G31900 iPEG 9 Wolff et al., 2011
AT5G50470 iPEG 9 Wolff et al., 2011
AT2G20160 iPEG 9 Wolff et al., 2011
AT4G29650 iMEG 9 Wolff et al., 2011
AT4G29570 iMEG 9 Wolff et al., 2011
AT4G29640  iMEG  9  Wolff et al., 2011
AT1G60970  iMEG  9  Wolff et al., 2011
AT4G26140  iMEG  9  Wolff et al., 2011
AT3G10590  iMEG  9  Wolff et al., 2011
AT3G10900  iMEG  9  Wolff et al., 2011
AT1G54280  iMEG  9  Wolff et al., 2011
AT5G08360  iMEG  9  Wolff et al., 2011
AT4G22050  iMEG  9  Wolff et al., 2011
AT5G06500  iMEG  9  Wolff et al., 2011
AT5G03020  iMEG  9  Wolff et al., 2011
AT3G26590  iMEG  9  Wolff et al., 2011
AT2G19400  iMEG  9  Wolff et al., 2011
AT3G21830  iMEG  9  Wolff et al., 2011
AT1G65330  iPEG  9  Wolff et al., 2011; Kohler et al., 2005
AT4G00540  iMEG  9  Wolff et al., 2011; Gehring et al., 2009
AT5G26650  iMEG  9  Wolff et al., 2011; Shirzadi et al., 2011
Supplementary Table 5.2: List of 72 imprinted genes that showed greater than 6 species in orthology clusters.

<table>
<thead>
<tr>
<th>GeneId</th>
<th>Status</th>
<th>Annotation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT5G53870*</td>
<td>iMEG</td>
<td>Early nodulin-like protein 1 (ENODL1)</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G31060</td>
<td>iMEG</td>
<td>Encodes a member of the DREB subfamily A-5 of ERF/AP2 transcription factor family</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT4G18650</td>
<td>iMEG</td>
<td>unknown protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT2G28380</td>
<td>iMEG</td>
<td>Encodes a cytoplasmic dsRNA-binding protein DRB2</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G03370</td>
<td>iMEG</td>
<td>Acylphosphatase family</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT3G22180</td>
<td>iMEG</td>
<td>DHHC-type zinc finger family protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>iMEG</td>
<td>Eukaryotic translation initiation factor 2 (eIF-2) family protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G22920</td>
<td>iMEG</td>
<td>CHY-type/CTCHY-type/RING-type Zinc finger protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G03280</td>
<td>iMEG</td>
<td>Involved in ethylene signal transduction</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G76250</td>
<td>iMEG</td>
<td>unknown protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT2G17990</td>
<td>iMEG</td>
<td>BEST Arabidopsis thaliana protein match is: kinecin-related (TAIR:AT5G66250.3)</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G77000</td>
<td>iMEG</td>
<td>AtSKP2:2 is a homolog of human SKP2, the human F-box protein that recruits E2F1</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G39955</td>
<td>iMEG</td>
<td>alpha/beta-Hydrolases superfamily protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G47770</td>
<td>iMEG</td>
<td>Encodes a protein with farnesyl diphosphate synthase activity</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G24460</td>
<td>iMEG</td>
<td>unknown protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G21790</td>
<td>iMEG</td>
<td>TRAM, LAG1 and CLN8 (TLC) lipid-sensing domain containing protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT5G53250</td>
<td>iMEG</td>
<td>arabinogalactan protein 22 (AGP22)</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G16760</td>
<td>iMEG</td>
<td>Encodes a medium to long-chain acyl-CoA oxidase</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT4G01840</td>
<td>iMEG</td>
<td>Encodes AtTPK5, a member of the Arabidopsis thaliana K+ channel family of AtTPK/KCO proteins</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>iMEG</td>
<td>Patched family protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G48910</td>
<td>iPEG</td>
<td>NADP binding, oxireductase activity</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>AT3G45090</td>
<td>iPEG P-loop containing nucleoside triphosphate hydrolases superfamily protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G63020</td>
<td>iPEG Encodes one of two alternative largest subunits of a putative plant-specific RNA polymerase IV</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G67820</td>
<td>iPEG Protein phosphatase 2C family protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G67830</td>
<td>iPEG Encodes a protein with α-fucosidase activity. The activity was assessed on 2'-fucosyl-lactitol</td>
<td>Wolff et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G60220</td>
<td>iPEG Encodes a deSUMOylating enzyme. In vitro it has both peptidase activity and isopeptidase activity</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G11460</td>
<td>iPEG Protein of unknown function (DUF581)</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT4G13460</td>
<td>iPEG Encodes a SU(VAR)3-9 homolog, a SET domain protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G70560</td>
<td>iPEG TAA1 is involved in the shade-induced production of indole-3-pyruvate (IPA), a precursor to IAA, a biologically active auxin</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT2G21450</td>
<td>iPEG chromatin remodeling 34 (CHR34)</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G60760</td>
<td>iPEG P-loop containing nucleoside triphosphate hydrolases superfamily protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G53150</td>
<td>iPEG DNAJ heat shock N-terminal domain-containing protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G49600</td>
<td>iPEG RNA-binding protein 47A (RBP47A)</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G37290</td>
<td>iPEG ARM repeat superfamily protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT4G15900</td>
<td>iPEG Mutations confer hypersensitivity to glucose and sucrose and augments sensitivity to cytokinin, ethylene, ABA and auxin</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G43780</td>
<td>iPEG sulfate adenyltransferase, ATP sulfurylase</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT3G16320</td>
<td>iPEG Subunit in the anaphase-promoting complex</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT2G40520</td>
<td>iPEG Nucleotidyltransferase family protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G47490</td>
<td>iPEG RNA-binding protein 47C (RBP47C)</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G42670</td>
<td>iPEG Agenet domain-containing protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT5G28300</td>
<td>iPEG Duplicated homeodomain-like superfamily protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G20910</td>
<td>iPEG ARID/BRIGHT DNA-binding domain-containing protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>AT1G18250</td>
<td>iPEG encodes a thaumatin-like protein</td>
<td>Gehring et al., 2011</td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Code</td>
<td>Description</td>
<td>Source</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>AT1G60970</td>
<td>iMEG</td>
<td>SNARE-like superfamily protein; intracellular protein transport, transport</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT3G26590</td>
<td>iMEG</td>
<td>MATE efflux family protein; antiporter activity, drug transmembrane transporter activity, transporter activity</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G73560</td>
<td>iMEG</td>
<td>Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G49290</td>
<td>iPEG</td>
<td>unknown protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G31900</td>
<td>iPEG</td>
<td>PICKLE RELATED 2 (PKR2)</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT3G62230</td>
<td>iPEG</td>
<td>Target promoter of the male germline-specific transcription factor DUO1</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G23320</td>
<td>iPEG</td>
<td>Encodes a protein with similarity to the TAA1 tryptophan aminotransferase involved in IAA biosynthesis</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G15140</td>
<td>iPEG</td>
<td>Galactose mutarotase-like superfamily protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G61330</td>
<td>iPEG</td>
<td>FBD, F-box and Leucine Rich Repeat domains containing protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G17770</td>
<td>iPEG</td>
<td>Encodes a SU(VAR)3-9 homolog, a SET domain protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT5G54650</td>
<td>iMEG</td>
<td>Encodes a protein with similarity to formins that is involved in cytokinesis</td>
<td>pre-2010</td>
</tr>
<tr>
<td>AT5G05440</td>
<td>iPEG</td>
<td>Encodes a member of the PYR (pyrabactin resistance )/PYL(PYR1-like)/RCAR (regulatory components of ABA receptor)</td>
<td>Unpublished data</td>
</tr>
<tr>
<td>AT1G07705</td>
<td>iPEG</td>
<td>NOT2 / NOT3 / NOT5 family; transcription regulator activity</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>iMEG</td>
<td>Embryo development ending in seed dormancy, pollen development, regulation of transcription, DNA-dependent</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G62660</td>
<td>iMEG</td>
<td>Glycosyl hydrolases family 32 protein</td>
<td>Hsieh et al., 2011</td>
</tr>
<tr>
<td>AT1G57800</td>
<td>iPEG</td>
<td>predicted to encode a protein with an N-terminal PHD domain and two RING domains surrounding an SRA domain</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT3G14205</td>
<td>iPEG</td>
<td>Phosphoinositide phosphatase family protein</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT2G32370</td>
<td>iPEG</td>
<td>Encodes a homeobox-leucine zipper family protein belonging to the HD-ZIP IV family</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G57820</td>
<td>iPEG</td>
<td>involved in centromere heterochromatinization</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>Accession</td>
<td>Type</td>
<td>Description</td>
<td>Reference</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>AT3G08040</td>
<td>iPEG</td>
<td>Encodes a member of the MATE (multidrug and toxin efflux family), expressed in roots but not shoots</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT1G59890</td>
<td>iPEG</td>
<td>SIN3-like 5 (SNL5)</td>
<td>Gehring et al., 2011</td>
</tr>
<tr>
<td>AT4G26140</td>
<td>iMEG</td>
<td>putative beta-galactosidase</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G55050</td>
<td>iMEG</td>
<td>unknown protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G59670</td>
<td>iMEG</td>
<td>Encodes glutathione transferase belonging to the tau class of GSTs</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT2G19400</td>
<td>iMEG</td>
<td>AGC (cAMP-dependent, cGMP-dependent and protein kinase C) kinase family protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G54280</td>
<td>iMEG</td>
<td>ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT1G62340</td>
<td>iMEG</td>
<td>Subtilisin-like serine protease required for epidermal surface formation in embryos and juvenile plants</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G11400*</td>
<td>iPEG</td>
<td>ARID/BRIGHT DNA-binding domain;ELM2 domain protein</td>
<td>Wolff et al., 2011</td>
</tr>
<tr>
<td>AT4G20800</td>
<td>iPEG</td>
<td>FAD-binding Berberine family protein</td>
<td>Wolff et al., 2011</td>
</tr>
</tbody>
</table>

* Imprinted genes marked with asterik (*) showed norMD score < 0.6 and were excluded from the codeML analysis. These two genes were included with the set of imprinted genes analyzed using McDonald-Kreitman test.
Supplementary Table 5.3: Genes from genome random dataset (100 genes) under lineage-specific and site-specific selection.

<table>
<thead>
<tr>
<th>Number</th>
<th>GenId</th>
<th>Lineage-specific Selection</th>
<th>Site-specific Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AT2G21540</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>2</td>
<td>AT3G16480</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>3</td>
<td>AT2G34090</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>4</td>
<td>AT4G28990</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>5</td>
<td>AT2G05310</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>6</td>
<td>AT4G04790</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>7</td>
<td>AT3G09760</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>8</td>
<td>AT1G76440</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>9</td>
<td>AT1G29820</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>10</td>
<td>AT2G40400</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>11</td>
<td>AT1G25375</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>12</td>
<td>AT5G46030</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>13</td>
<td>AT2G01220</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>14</td>
<td>AT4G13260</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>15</td>
<td>AT5G26910</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>16</td>
<td>AT1G21790</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>17</td>
<td>AT1G77810</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>18</td>
<td>AT5G64460</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>19</td>
<td>AT2G34480</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>20</td>
<td>AT2G34430</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>21</td>
<td>AT4G21500</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>22</td>
<td>AT5G60580</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>23</td>
<td>AT3G50690</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>24</td>
<td>AT4G03100</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>25</td>
<td>AT4G23820</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>26</td>
<td>AT4G10440</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>27</td>
<td>AT5G49510</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>28</td>
<td>AT5G56360</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>29</td>
<td>AT2G13800</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>30</td>
<td>AT5G59860</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>31</td>
<td>AT5G62690</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>32</td>
<td>AT5G19980</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>33</td>
<td>AT3G26630</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>34</td>
<td>AT1G78190</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>35</td>
<td>AT3G55510</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>36</td>
<td>AT2G40840</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>37</td>
<td>AT2G17265</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>38</td>
<td>AT5G43740</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>39</td>
<td>AT4G16950</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>40</td>
<td>AT4G19380</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>41</td>
<td>AT2G02710</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>42</td>
<td>AT3G50360</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>AT5G04720</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>---</td>
<td>---------</td>
</tr>
<tr>
<td>43</td>
<td>AT3G14430</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>44</td>
<td>AT5G18810</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>45</td>
<td>AT1G04645</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>46</td>
<td>AT1G17950</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>47</td>
<td>AT2G35010</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>48</td>
<td>AT1G79040</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>49</td>
<td>AT2G30160</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>50</td>
<td>AT3G46630</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>51</td>
<td>AT3G09085</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>52</td>
<td>AT5G02030</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>53</td>
<td>AT5G20630</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>54</td>
<td>AT5G56140</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>55</td>
<td>AT3G45860</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>56</td>
<td>AT1G49310</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>57</td>
<td>AT4G24580</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>58</td>
<td>AT2G07020</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>59</td>
<td>AT5G51030</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>60</td>
<td>AT3G05560</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>61</td>
<td>AT2G40950</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>62</td>
<td>AT2G31920</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>63</td>
<td>AT1G76070</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>64</td>
<td>AT1G80500</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>65</td>
<td>AT4G02730</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>66</td>
<td>AT1G56460</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>67</td>
<td>AT2G30240</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>68</td>
<td>AT3G20620</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>69</td>
<td>AT1G25510</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>70</td>
<td>AT3G03790</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>71</td>
<td>AT3G27310</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>72</td>
<td>AT3G46510</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>73</td>
<td>AT4G17910</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>74</td>
<td>AT2G31650</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>75</td>
<td>AT1G60490</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>76</td>
<td>AT5G35730</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>77</td>
<td>AT2G20470</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>78</td>
<td>AT5G28080</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>79</td>
<td>AT4G31320</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>80</td>
<td>AT5G10150</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>81</td>
<td>AT5G62580</td>
<td></td>
<td>Present</td>
</tr>
<tr>
<td>82</td>
<td>AT3G60550</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>83</td>
<td>AT1G55610</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>84</td>
<td>AT1G27385</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>85</td>
<td>AT1G49880</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>86</td>
<td>AT2G24590</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>87</td>
<td>AT4G24660</td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td></td>
<td>AT1G55870</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>90</td>
<td>AT4G05420</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>91</td>
<td>AT3G60780</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>92</td>
<td>AT5G60820</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>93</td>
<td>AT5G01015</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>94</td>
<td>AT3G50830</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>95</td>
<td>AT3G19650</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>96</td>
<td>AT1G02520</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>97</td>
<td>AT4G24670</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>98</td>
<td>AT1G55920</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>99</td>
<td>AT5G40390</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>100</td>
<td>AT4G33720</td>
<td>Absent</td>
<td>Absent</td>
</tr>
</tbody>
</table>
Supplementary Table 5.4: Genes from endosperm-specific random dataset (100 genes) under lineage-specific and site-specific selection.

<table>
<thead>
<tr>
<th>Number</th>
<th>GeneId</th>
<th>Lineage-specific selection</th>
<th>Site-specific selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AT5G05850</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>2</td>
<td>AT4G18050</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>3</td>
<td>AT4G39910</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>4</td>
<td>AT4G23890</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>5</td>
<td>AT5G06970</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>6</td>
<td>AT2G39730</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>7</td>
<td>AT1G27120</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>8</td>
<td>AT5G59100</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>9</td>
<td>AT5G46700</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>10</td>
<td>AT2G35820</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>11</td>
<td>AT5G57180</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>12</td>
<td>AT1G45474</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>13</td>
<td>AT3G21400</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>14</td>
<td>AT1G71020</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>15</td>
<td>AT4G29380</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>16</td>
<td>AT3G49600</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>17</td>
<td>AT5G52160</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>18</td>
<td>AT1G77310</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>19</td>
<td>AT2G14520</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>20</td>
<td>AT5G51030</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>21</td>
<td>AT1G26230</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>22</td>
<td>AT2G38720</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>23</td>
<td>AT5G59190</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>24</td>
<td>AT1G64290</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>25</td>
<td>AT1G63120</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>26</td>
<td>AT1G10680</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>27</td>
<td>AT1G35720</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>28</td>
<td>AT1G09380</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>29</td>
<td>AT1G79960</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>30</td>
<td>AT2G45530</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>31</td>
<td>AT3G25160</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>32</td>
<td>AT5G08280</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>33</td>
<td>AT1G50900</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>34</td>
<td>AT3G02730</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>35</td>
<td>AT1G61320</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>36</td>
<td>AT4G19100</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>37</td>
<td>AT2G20840</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>38</td>
<td>AT2G19500</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>39</td>
<td>AT1G48260</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>40</td>
<td>AT3G23510</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>41</td>
<td>AT1G15180</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>AT5G67370</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>43</td>
<td>AT4G02710</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>44</td>
<td>AT5G16370</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>45</td>
<td>AT3G52780</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>46</td>
<td>AT1G51210</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>47</td>
<td>AT2G03420</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>48</td>
<td>AT5G39290</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>49</td>
<td>AT1G70560</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>50</td>
<td>AT2G25630</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>51</td>
<td>AT5G44840</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>52</td>
<td>AT2G46430</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>53</td>
<td>AT5G20710</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>54</td>
<td>AT3G27720</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>55</td>
<td>AT5G28500</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>56</td>
<td>AT1G55370</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>57</td>
<td>AT4G01120</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>58</td>
<td>AT3G06860</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>59</td>
<td>AT5G12460</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>60</td>
<td>AT5G42310</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>61</td>
<td>AT2G19900</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>62</td>
<td>AT1G49810</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>63</td>
<td>AT4G38780</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>64</td>
<td>AT2G38920</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>65</td>
<td>AT5G38900</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>66</td>
<td>AT1G03630</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>67</td>
<td>AT4G29050</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>68</td>
<td>AT3G55800</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>69</td>
<td>AT5G12960</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>70</td>
<td>AT5G58830</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>71</td>
<td>AT1G01320</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>72</td>
<td>AT5G13210</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>73</td>
<td>AT5G48000</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>74</td>
<td>AT1G08980</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>75</td>
<td>AT3G12930</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>76</td>
<td>AT5G03840</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>77</td>
<td>AT5G09500</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>78</td>
<td>AT2G39420</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>79</td>
<td>AT3G27710</td>
<td>Present</td>
<td>Absent</td>
</tr>
<tr>
<td>80</td>
<td>AT3G48950</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>81</td>
<td>AT4G22030</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>82</td>
<td>AT2G02120</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>83</td>
<td>AT5G18310</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>84</td>
<td>AT5G38710</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>85</td>
<td>AT2G26870</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>86</td>
<td>AT1G78700</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>87</td>
<td>AT5G42070</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td></td>
<td>AT5G47730</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>---</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>88</td>
<td>AT3G61870</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>89</td>
<td>AT5G13200</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>90</td>
<td>AT4G22830</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>91</td>
<td>AT5G17540</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>92</td>
<td>AT2G19760</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>93</td>
<td>AT5G53520</td>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>94</td>
<td>AT3G12960</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>95</td>
<td>AT2G31760</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>96</td>
<td>AT2G31890</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>97</td>
<td>AT5G63770</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>98</td>
<td>AT5G62460</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>99</td>
<td>AT3G27770</td>
<td>Absent</td>
<td>Present</td>
</tr>
</tbody>
</table>
Supplementary Table 5.5: Percentage of fixation of lineage-specific positively selected sites in Ath 80 population data identified by codeml.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Sites</th>
<th>Posterior probability</th>
<th>Ath amino acid</th>
<th>% of conservation of sites in Ath 80 accessions</th>
<th>Aly amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT2G28380</td>
<td>272</td>
<td>0.982</td>
<td>N</td>
<td>100</td>
<td>A</td>
</tr>
<tr>
<td>361</td>
<td>0.589</td>
<td>K</td>
<td>100</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>367</td>
<td>0.744</td>
<td>G</td>
<td>100</td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>368</td>
<td>0.66</td>
<td>E</td>
<td>98.75</td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>370</td>
<td>0.675</td>
<td>R</td>
<td>100</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>372</td>
<td>0.739</td>
<td>S</td>
<td>100</td>
<td></td>
<td>V</td>
</tr>
<tr>
<td>374</td>
<td>0.904</td>
<td>V</td>
<td>100</td>
<td></td>
<td>Q</td>
</tr>
<tr>
<td>402</td>
<td>0.864</td>
<td>E</td>
<td>100</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>407</td>
<td>0.611</td>
<td>K</td>
<td>100</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>414</td>
<td>0.691</td>
<td>F</td>
<td>100</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>53</td>
<td>0.508</td>
<td>E</td>
<td>100</td>
<td>A</td>
</tr>
<tr>
<td>108</td>
<td>0.951</td>
<td>D</td>
<td>100</td>
<td></td>
<td>K</td>
</tr>
<tr>
<td>136</td>
<td>0.583</td>
<td>T</td>
<td>100</td>
<td></td>
<td>K</td>
</tr>
<tr>
<td>178</td>
<td>0.962</td>
<td>E</td>
<td>100</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>181</td>
<td>0.522</td>
<td>P</td>
<td>100</td>
<td></td>
<td>S</td>
</tr>
<tr>
<td>187</td>
<td>0.519</td>
<td>K</td>
<td>100</td>
<td></td>
<td>D</td>
</tr>
<tr>
<td>193</td>
<td>0.533</td>
<td>Q</td>
<td>100</td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>225</td>
<td>0.974</td>
<td>E</td>
<td>100</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>235</td>
<td>0.51</td>
<td>T</td>
<td>100</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>240</td>
<td>0.908</td>
<td>P</td>
<td>100</td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>247</td>
<td>0.62</td>
<td>S</td>
<td>100</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>273</td>
<td>0.592</td>
<td>A</td>
<td>100</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>276</td>
<td>0.643</td>
<td>Q</td>
<td>100</td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>291</td>
<td>0.871</td>
<td>Q</td>
<td>100</td>
<td></td>
<td>L</td>
</tr>
<tr>
<td>475</td>
<td>0.612</td>
<td>I</td>
<td>100</td>
<td></td>
<td>L</td>
</tr>
<tr>
<td>581</td>
<td>0.966</td>
<td>L</td>
<td>100</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>AT1G76820</td>
<td>170</td>
<td>0.85</td>
<td>S</td>
<td>100</td>
<td>NA</td>
</tr>
<tr>
<td>172</td>
<td>0.532</td>
<td>R</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>173</td>
<td>0.774</td>
<td>S</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>176</td>
<td>0.881</td>
<td>H</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>186</td>
<td>0.873</td>
<td>L</td>
<td>92.5</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>227</td>
<td>0.715</td>
<td>N</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>234</td>
<td>0.874</td>
<td>R</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>238</td>
<td>0.884</td>
<td>R</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>239</td>
<td>0.89</td>
<td>L</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>240</td>
<td>0.954</td>
<td>L</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>241</td>
<td>0.937</td>
<td>Q</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>250</td>
<td>0.917</td>
<td>R</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>342</td>
<td>0.978</td>
<td>A</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>422</td>
<td>0.759</td>
<td>G</td>
<td>100</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Position</td>
<td>Similarity</td>
<td>Identity</td>
<td>Database ID</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>------------</td>
<td>----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>AT1G42470</td>
<td>42</td>
<td>0.538</td>
<td>63.75</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>338</td>
<td>0.566</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>622</td>
<td>0.511</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>0.547</td>
<td>97.5</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1187</td>
<td>0.53</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1210</td>
<td>0.536</td>
<td>98.75</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1221</td>
<td>0.536</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1240</td>
<td>0.975</td>
<td>100</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>AT1G48910</td>
<td>153</td>
<td>0.992</td>
<td>96.25</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>362</td>
<td>0.503</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>AT1G63020</td>
<td>77</td>
<td>0.948</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>201</td>
<td>0.751</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>932</td>
<td>0.944</td>
<td>1.25</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1177</td>
<td>0.723</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1287</td>
<td>0.56</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>AT5G53150</td>
<td>49</td>
<td>0.512</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>0.722</td>
<td>100</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>115</td>
<td>0.816</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.509</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>439</td>
<td>0.675</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td></td>
<td>442</td>
<td>0.794</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td></td>
<td>509</td>
<td>0.539</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>526</td>
<td>0.501</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>624</td>
<td>0.535</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AT1G47490</td>
<td>57</td>
<td>0.53</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>0.994</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>0.778</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>0.851</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>139</td>
<td>0.825</td>
<td>100</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td></td>
<td>162</td>
<td>0.828</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>205</td>
<td>0.855</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>301</td>
<td>0.83</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td></td>
<td>325</td>
<td>0.996</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AT2G40520</td>
<td>5</td>
<td>0.638</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>0.74</td>
<td>100</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>0.826</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>0.606</td>
<td>S</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>0.712</td>
<td>N</td>
<td>100</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>0.792</td>
<td>L</td>
<td>100</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>0.777</td>
<td>Y</td>
<td>100</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>0.746</td>
<td>M</td>
<td>98.75</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.562</td>
<td>H</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>0.704</td>
<td>A</td>
<td>100</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>0.716</td>
<td>S</td>
<td>100</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>0.656</td>
<td>N</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td>0.801</td>
<td>L</td>
<td>100</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>0.624</td>
<td>G</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>0.79</td>
<td>V</td>
<td>100</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>0.787</td>
<td>K</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>145</td>
<td>0.745</td>
<td>D</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>0.786</td>
<td>F</td>
<td>100</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>0.955</td>
<td>K</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>184</td>
<td>0.687</td>
<td>V</td>
<td>100</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>233</td>
<td>0.724</td>
<td>I</td>
<td>100</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>0.697</td>
<td>N</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>246</td>
<td>0.661</td>
<td>V</td>
<td>100</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>285</td>
<td>0.99</td>
<td>E</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>287</td>
<td>0.769</td>
<td>G</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>330</td>
<td>0.592</td>
<td>M</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>333</td>
<td>0.677</td>
<td>V</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>356</td>
<td>0.71</td>
<td>N</td>
<td>100</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>357</td>
<td>0.992</td>
<td>V</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>381</td>
<td>0.975</td>
<td>A</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>386</td>
<td>0.926</td>
<td>Q</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>395</td>
<td>0.817</td>
<td>R</td>
<td>100</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>423</td>
<td>0.636</td>
<td>A</td>
<td>98.75</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>427</td>
<td>0.801</td>
<td>F</td>
<td>100</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>431</td>
<td>0.533</td>
<td>R</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>432</td>
<td>0.748</td>
<td>V</td>
<td>92.5</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>442</td>
<td>0.65</td>
<td>E</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>478</td>
<td>0.793</td>
<td>C</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>479</td>
<td>0.817</td>
<td>S</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>AT5G28300</td>
<td>72</td>
<td>0.506</td>
<td>I</td>
<td>100</td>
<td>D</td>
</tr>
<tr>
<td>575</td>
<td>0.869</td>
<td>R</td>
<td>85</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>576</td>
<td>0.632</td>
<td>D</td>
<td>100</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>578</td>
<td>0.562</td>
<td>D</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>579</td>
<td>0.736</td>
<td>T</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>581</td>
<td>0.551</td>
<td>P</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>AT1G20190</td>
<td>23</td>
<td>0.514</td>
<td>E</td>
<td>100</td>
<td>P</td>
</tr>
<tr>
<td>84</td>
<td>0.677</td>
<td>N</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>0.635</td>
<td>L</td>
<td>100</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>0.969</td>
<td>A</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td></td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>0.88</td>
<td>L</td>
<td>100</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>0.664</td>
<td>R</td>
<td>100</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>0.667</td>
<td>S</td>
<td>100</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>139</td>
<td>0.962</td>
<td>I</td>
<td>100</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>0.914</td>
<td>V</td>
<td>78.75</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>0.697</td>
<td>E</td>
<td>100</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>0.565</td>
<td>N</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>0.973</td>
<td>Y</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>183</td>
<td>0.518</td>
<td>N</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>195</td>
<td>0.995</td>
<td>C</td>
<td>100</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>0.969</td>
<td>N</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>234</td>
<td>0.805</td>
<td>S</td>
<td>100</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>0.807</td>
<td>A</td>
<td>97.5</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>257</td>
<td>0.702</td>
<td>A</td>
<td>100</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>269</td>
<td>0.542</td>
<td>H</td>
<td>100</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>270</td>
<td>0.637</td>
<td>K</td>
<td>100</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>308</td>
<td>0.996</td>
<td>V</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>317</td>
<td>0.799</td>
<td>K</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>322</td>
<td>0.881</td>
<td>S</td>
<td>100</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>335</td>
<td>0.72</td>
<td>K</td>
<td>100</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>339</td>
<td>0.649</td>
<td>I</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>346</td>
<td>0.969</td>
<td>K</td>
<td>100</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>347</td>
<td>0.829</td>
<td>V</td>
<td>100</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>368</td>
<td>0.716</td>
<td>I</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>372</td>
<td>0.85</td>
<td>S</td>
<td>100</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>384</td>
<td>0.78</td>
<td>M</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>386</td>
<td>0.813</td>
<td>M</td>
<td>100</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>392</td>
<td>0.815</td>
<td>I</td>
<td>100</td>
<td>V</td>
<td></td>
</tr>
</tbody>
</table>

**AT4G31900**

<p>| 3  | 0.767| N  | 100| S  |
| 5  | 0.835| L  | 100| V  |
| 6  | 0.78| Q  | 100| H  |
| 10 | 0.598| R  | 100| P  |
| 19 | 0.716| I  | 16.25| T  |
| 22 | 0.992| K  | 100| D  |
| 26 | 0.839| Y  | 100| D  |
| 27 | 0.787| I  | 100| F  |
| 28 | 0.722| R  | 100| -  |
| 31 | 0.989| Q  | 100| V  |
| 39 | 0.753| N  | 100| K  |
| 40 | 0.856| Q  | 100| E  |
| 41 | 0.74| D  | 100| E  |
| 55 | 0.866| W  | 100| L  |
| 62 | 0.733| N  | 100| D  |
| 73 | 0.848| L  | 100| Q  |
| 75 | 0.856| V  | 100| F  |
| 108| 0.769| H  | 100| R  |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>0.865</td>
<td>L</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>0.971</td>
<td>T</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>0.843</td>
<td>R</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131</td>
<td>0.765</td>
<td>A</td>
<td>100</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>0.794</td>
<td>H</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>0.832</td>
<td>K</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>0.608</td>
<td>E</td>
<td>100</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>157</td>
<td>0.655</td>
<td>E</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>0.994</td>
<td>R</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>0.984</td>
<td>N</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>0.731</td>
<td>D</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>195</td>
<td>0.607</td>
<td>S</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>0.566</td>
<td>E</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>216</td>
<td>0.814</td>
<td>L</td>
<td>100</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>0.991</td>
<td>K</td>
<td>100</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>284</td>
<td>0.87</td>
<td>I</td>
<td>100</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>306</td>
<td>0.976</td>
<td>D</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>321</td>
<td>0.839</td>
<td>S</td>
<td>100</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>322</td>
<td>0.661</td>
<td>E</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>326</td>
<td>0.604</td>
<td>S</td>
<td>76.25</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>331</td>
<td>0.774</td>
<td>L</td>
<td>100</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>334</td>
<td>0.866</td>
<td>T</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>338</td>
<td>0.796</td>
<td>V</td>
<td>100</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>339</td>
<td>0.842</td>
<td>H</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>341</td>
<td>0.984</td>
<td>G</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>342</td>
<td>0.599</td>
<td>I</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346</td>
<td>0.993</td>
<td>S</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>355</td>
<td>0.794</td>
<td>I</td>
<td>100</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>364</td>
<td>0.826</td>
<td>Q</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>365</td>
<td>0.963</td>
<td>K</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>373</td>
<td>0.757</td>
<td>S</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>380</td>
<td>0.997</td>
<td>I</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>393</td>
<td>0.738</td>
<td>N</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>397</td>
<td>0.638</td>
<td>A</td>
<td>100</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>418</td>
<td>0.824</td>
<td>K</td>
<td>100</td>
<td>Q</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>446</td>
<td>0.8</td>
<td>V</td>
<td>100</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>458</td>
<td>0.833</td>
<td>M</td>
<td>100</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>482</td>
<td>0.777</td>
<td>D</td>
<td>100</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>491</td>
<td>0.817</td>
<td>K</td>
<td>97.5</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>494</td>
<td>0.822</td>
<td>Q</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>503</td>
<td>0.993</td>
<td>P</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>505</td>
<td>0.829</td>
<td>F</td>
<td>100</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>554</td>
<td>0.852</td>
<td>Y</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>556</td>
<td>0.821</td>
<td>L</td>
<td>100</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>560</td>
<td>0.871</td>
<td>F</td>
<td>100</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>574</td>
<td>0.853</td>
<td>I</td>
<td>100</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Score</td>
<td>Type</td>
<td>Probability</td>
<td>Score</td>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
<td>-------------</td>
<td>-------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT1G61330</td>
<td>15</td>
<td></td>
<td>0.898</td>
<td>L</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>575</td>
<td>0.722</td>
<td>S</td>
<td>100</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>577</td>
<td>0.837</td>
<td>P</td>
<td>100</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>599</td>
<td>0.848</td>
<td>S</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>634</td>
<td>0.809</td>
<td>V</td>
<td>100</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>651</td>
<td>0.96</td>
<td>K</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>659</td>
<td>0.627</td>
<td>M</td>
<td>100</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>664</td>
<td>0.798</td>
<td>N</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>676</td>
<td>0.778</td>
<td>Q</td>
<td>100</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>679</td>
<td>0.754</td>
<td>C</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>704</td>
<td>0.797</td>
<td>R</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>717</td>
<td>0.827</td>
<td>Q</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>722</td>
<td>0.739</td>
<td>N</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>727</td>
<td>0.862</td>
<td>V</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>737</td>
<td>0.805</td>
<td>T</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>738</td>
<td>0.778</td>
<td>D</td>
<td>100</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>747</td>
<td>0.62</td>
<td>S</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>751</td>
<td>0.815</td>
<td>V</td>
<td>100</td>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>754</td>
<td>0.557</td>
<td>E</td>
<td>100</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>883</td>
<td>0.644</td>
<td>Y</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>885</td>
<td>0.752</td>
<td>M</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>891</td>
<td>0.59</td>
<td>E</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>895</td>
<td>0.75</td>
<td>D</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>916</td>
<td>0.8</td>
<td>N</td>
<td>100</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>919</td>
<td>0.813</td>
<td>Y</td>
<td>100</td>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>942</td>
<td>0.544</td>
<td>T</td>
<td>100</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>949</td>
<td>0.996</td>
<td>K</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>966</td>
<td>0.792</td>
<td>S</td>
<td>100</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>980</td>
<td>0.638</td>
<td>V</td>
<td>100</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>989</td>
<td>0.81</td>
<td>N</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>998</td>
<td>0.796</td>
<td>N</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.864</td>
<td>V</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1008</td>
<td>0.745</td>
<td>N</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1031</td>
<td>0.961</td>
<td>R</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1033</td>
<td>0.981</td>
<td>V</td>
<td>100</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1095</td>
<td>0.743</td>
<td>K</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1097</td>
<td>0.726</td>
<td>E</td>
<td>100</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1099</td>
<td>0.695</td>
<td>K</td>
<td>100</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1105</td>
<td>0.678</td>
<td>D</td>
<td>100</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1120</td>
<td>0.757</td>
<td>K</td>
<td>94.9</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1131</td>
<td>0.661</td>
<td>G</td>
<td>98.75</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1140</td>
<td>0.741</td>
<td>D</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1152</td>
<td>0.794</td>
<td>Q</td>
<td>100</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1156</td>
<td>0.782</td>
<td>D</td>
<td>100</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1164</td>
<td>0.84</td>
<td>P</td>
<td>100</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1181</td>
<td>0.796</td>
<td>R</td>
<td>100</td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**AT1G61330**
<p>| AT1G17770 | 34  | 0.676 | V    | 95   | -     |
| AT1G49600 | 105 | 0.564 | S    | 100  | G     |
| AT1G5890 | 295 | 0.789 | P    | 100  | -     |
| AT1G57820 | 218 | 0.924 | T    | 13.75| G     |
| AT1G67820 | 239 | 0.522 | A    | 100  | R     |
| AT1G55050 | 26  | 0.553 | I    | 100  | V     |
| AT1G54280 | 965 | 0.971 | C    | 100  | A     |
| AT1G57820 | 442 | 0.947 | I    | 100  | L     |
| AT1G59890 | 565 | 0.831 | R    | 97.5 | Q     |
| AT1G55050 | 592 | 0.94  | A    | 100  | M     |
| AT1G57820 | 605 | 0.6   | T    | 50   | V     |
| AT1G55050 | 622 | 0.665 | E    | 100  | S     |
| AT1G67820 | 404 | 0.982 | P    | 100  | K     |
| AT1G55050 | 410 | 0.975 | S    | 100  | R     |
| AT1G55050 | 26  | 0.553 | I    | 100  | V     |
| AT1G55050 | 40   | 0.714 | M    | 100  | I     |
| AT1G55050 | 47   | 0.503 | E    | 100  | A     |
| AT1G55050 | 54   | 0.728 | E    | 100  | D     |
| AT1G55050 | 173  | 0.957 | G    | 100  | E     |
| AT1G55050 | 179  | 0.754 | T    | 100  | I     |
| AT1G55050 | 183  | 0.838 | Y    | 100  | S     |
| AT1G55050 | 189  | 0.877 | T    | 100  | R     |
| AT1G55050 | 246  | 0.722 | K    | 100  | E     |
| AT1G55050 | 272  | 0.733 | K    | 100  | E     |
| AT1G55050 | 275  | 0.711 | D    | 100  | N     |
| AT1G55050 | 277  | 0.734 | E    | 100  | K     |
| AT1G55050 | 294  | 0.615 | N    | 100  | S     |
| AT1G55050 | 306  | 0.842 | S    | 100  | T     |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>311</td>
<td>0.894</td>
<td>V</td>
<td>100</td>
</tr>
<tr>
<td>337</td>
<td>0.899</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>341</td>
<td>0.539</td>
<td>Q</td>
<td>2.5</td>
</tr>
<tr>
<td>348</td>
<td>0.769</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>363</td>
<td>0.988</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>368</td>
<td>0.696</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td>392</td>
<td>0.886</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>396</td>
<td>0.953</td>
<td>E</td>
<td>83.75</td>
</tr>
<tr>
<td>401</td>
<td>0.825</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>433</td>
<td>0.832</td>
<td>F</td>
<td>100</td>
</tr>
<tr>
<td>449</td>
<td>0.693</td>
<td>S</td>
<td>91.25</td>
</tr>
<tr>
<td>450</td>
<td>0.766</td>
<td>L</td>
<td>85</td>
</tr>
<tr>
<td>451</td>
<td>0.976</td>
<td>A</td>
<td>100</td>
</tr>
<tr>
<td>453</td>
<td>0.526</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>470</td>
<td>0.681</td>
<td>F</td>
<td>100</td>
</tr>
<tr>
<td>484</td>
<td>0.749</td>
<td>V</td>
<td>98.75</td>
</tr>
<tr>
<td>504</td>
<td>0.55</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>566</td>
<td>0.607</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>567</td>
<td>0.509</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>579</td>
<td>0.848</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>580</td>
<td>0.974</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>581</td>
<td>0.795</td>
<td>L</td>
<td>98.75</td>
</tr>
<tr>
<td>582</td>
<td>0.986</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td>583</td>
<td>0.992</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>584</td>
<td>0.952</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td>608</td>
<td>0.501</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>611</td>
<td>0.908</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>658</td>
<td>0.666</td>
<td>H</td>
<td>100</td>
</tr>
<tr>
<td>659</td>
<td>0.912</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>661</td>
<td>0.831</td>
<td>S</td>
<td>92.5</td>
</tr>
<tr>
<td>662</td>
<td>0.931</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>665</td>
<td>0.734</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>669</td>
<td>0.85</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>672</td>
<td>0.844</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>673</td>
<td>0.706</td>
<td>R</td>
<td>92.5</td>
</tr>
<tr>
<td>675</td>
<td>0.797</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>676</td>
<td>0.628</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>677</td>
<td>0.748</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>683</td>
<td>0.75</td>
<td>D</td>
<td>82.5</td>
</tr>
<tr>
<td>684</td>
<td>0.863</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>695</td>
<td>0.513</td>
<td>N</td>
<td>100</td>
</tr>
<tr>
<td>696</td>
<td>0.803</td>
<td>G</td>
<td>92.5</td>
</tr>
<tr>
<td>697</td>
<td>0.906</td>
<td>L</td>
<td>91.25</td>
</tr>
<tr>
<td>698</td>
<td>0.971</td>
<td>Y</td>
<td>85</td>
</tr>
<tr>
<td>699</td>
<td>0.956</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>700</td>
<td>0.937</td>
<td>S</td>
<td>98.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>701</td>
<td>0.924</td>
<td>K</td>
<td>93.6</td>
</tr>
<tr>
<td>702</td>
<td>0.982</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>703</td>
<td>0.926</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>704</td>
<td>0.561</td>
<td>G</td>
<td>100</td>
</tr>
<tr>
<td>705</td>
<td>0.551</td>
<td>A</td>
<td>100</td>
</tr>
<tr>
<td>706</td>
<td>0.978</td>
<td>F</td>
<td>92.5</td>
</tr>
<tr>
<td>707</td>
<td>0.893</td>
<td>H</td>
<td>100</td>
</tr>
<tr>
<td>709</td>
<td>0.585</td>
<td>H</td>
<td>92.5</td>
</tr>
<tr>
<td>710</td>
<td>0.633</td>
<td>S</td>
<td>92.5</td>
</tr>
<tr>
<td>711</td>
<td>0.883</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>712</td>
<td>0.696</td>
<td>T</td>
<td>92.5</td>
</tr>
<tr>
<td>716</td>
<td>0.94</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>717</td>
<td>0.973</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>718</td>
<td>0.716</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>719</td>
<td>0.938</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>720</td>
<td>0.799</td>
<td>N</td>
<td>100</td>
</tr>
<tr>
<td>721</td>
<td>0.518</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td>722</td>
<td>0.912</td>
<td>L</td>
<td>92.5</td>
</tr>
<tr>
<td>723</td>
<td>0.66</td>
<td>C</td>
<td>100</td>
</tr>
<tr>
<td>724</td>
<td>0.957</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>726</td>
<td>0.979</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>729</td>
<td>1</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>731</td>
<td>0.734</td>
<td>K</td>
<td>98.75</td>
</tr>
<tr>
<td>732</td>
<td>0.934</td>
<td>D</td>
<td>92.5</td>
</tr>
<tr>
<td>735</td>
<td>0.861</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>739</td>
<td>0.928</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>740</td>
<td>0.83</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>741</td>
<td>0.673</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>742</td>
<td>0.848</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>744</td>
<td>0.533</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>745</td>
<td>0.918</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>748</td>
<td>0.756</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>749</td>
<td>0.92</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>750</td>
<td>0.871</td>
<td>A</td>
<td>92.5</td>
</tr>
<tr>
<td>752</td>
<td>0.886</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>755</td>
<td>0.928</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>756</td>
<td>0.896</td>
<td>P</td>
<td>98.75</td>
</tr>
<tr>
<td>758</td>
<td>0.905</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td>759</td>
<td>0.931</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>760</td>
<td>0.879</td>
<td>H</td>
<td>100</td>
</tr>
<tr>
<td>761</td>
<td>0.563</td>
<td>G</td>
<td>100</td>
</tr>
<tr>
<td>762</td>
<td>0.743</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>763</td>
<td>0.942</td>
<td>T</td>
<td>100</td>
</tr>
<tr>
<td>764</td>
<td>0.951</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>765</td>
<td>0.807</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>766</td>
<td>0.808</td>
<td>R</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>768</td>
<td>0.891</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>769</td>
<td>0.889</td>
<td>L</td>
<td>100</td>
</tr>
<tr>
<td>770</td>
<td>0.925</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>771</td>
<td>0.688</td>
<td>Q</td>
<td>100</td>
</tr>
<tr>
<td>780</td>
<td>0.521</td>
<td>A</td>
<td>100</td>
</tr>
<tr>
<td>806</td>
<td>0.86</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>813</td>
<td>0.812</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>846</td>
<td>0.632</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>854</td>
<td>0.961</td>
<td>T</td>
<td>98.75</td>
</tr>
<tr>
<td>855</td>
<td>0.852</td>
<td>S</td>
<td>98.75</td>
</tr>
<tr>
<td>856</td>
<td>0.7</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>857</td>
<td>0.933</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>858</td>
<td>0.623</td>
<td>S</td>
<td>100</td>
</tr>
<tr>
<td>859</td>
<td>0.782</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>861</td>
<td>0.828</td>
<td>I</td>
<td>100</td>
</tr>
<tr>
<td>862</td>
<td>0.502</td>
<td>K</td>
<td>100</td>
</tr>
<tr>
<td>865</td>
<td>0.848</td>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td>866</td>
<td>0.778</td>
<td>P</td>
<td>100</td>
</tr>
<tr>
<td>870</td>
<td>0.585</td>
<td>V</td>
<td>100</td>
</tr>
</tbody>
</table>
Supplementary Table 5.6: Mcdonald-Kreitman calculations for *A. thaliana* imprinted genes.

<table>
<thead>
<tr>
<th>GeneId</th>
<th>Ds</th>
<th>Dn</th>
<th>Ps</th>
<th>Pn</th>
<th>NI</th>
<th>Alfa</th>
<th>Fisher-test P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01530</td>
<td>31</td>
<td>63</td>
<td>9</td>
<td>22</td>
<td>1.203</td>
<td>-0.203</td>
<td>0.825</td>
</tr>
<tr>
<td>AT1G02580</td>
<td>24</td>
<td>64</td>
<td>11</td>
<td>14</td>
<td>0.477</td>
<td>0.523</td>
<td>0.142</td>
</tr>
<tr>
<td>AT1G07690</td>
<td>12</td>
<td>31</td>
<td>2</td>
<td>8</td>
<td>1.548</td>
<td>-0.548</td>
<td>0.713</td>
</tr>
<tr>
<td>AT1G07705</td>
<td>26</td>
<td>20</td>
<td>12</td>
<td>13</td>
<td>1.408</td>
<td>-0.408</td>
<td>0.619</td>
</tr>
<tr>
<td>AT1G11810</td>
<td>23</td>
<td>67</td>
<td>2</td>
<td>10</td>
<td>1.716</td>
<td>-0.716</td>
<td>0.725</td>
</tr>
<tr>
<td>AT1G12180</td>
<td>9</td>
<td>35</td>
<td>3</td>
<td>11</td>
<td>0.943</td>
<td>0.057</td>
<td>1</td>
</tr>
<tr>
<td>AT1G18250</td>
<td>15</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>0.625</td>
<td>0.375</td>
<td>0.682</td>
</tr>
<tr>
<td>AT1G19160</td>
<td>32</td>
<td>82</td>
<td>3</td>
<td>13</td>
<td>1.691</td>
<td>-0.691</td>
<td>0.556</td>
</tr>
<tr>
<td>AT1G20680</td>
<td>24</td>
<td>55</td>
<td>8</td>
<td>13</td>
<td>0.709</td>
<td>0.291</td>
<td>0.599</td>
</tr>
<tr>
<td>AT1G20910</td>
<td>21</td>
<td>23</td>
<td>5</td>
<td>5</td>
<td>0.913</td>
<td>0.087</td>
<td>1</td>
</tr>
<tr>
<td>AT1G21790</td>
<td>24</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>0.706</td>
<td>0.294</td>
<td>0.759</td>
</tr>
<tr>
<td>AT1G23320</td>
<td>34</td>
<td>22</td>
<td>2</td>
<td>19</td>
<td>14.682</td>
<td>-13.682</td>
<td>0.000067</td>
</tr>
<tr>
<td>AT1G42470</td>
<td>63</td>
<td>37</td>
<td>60</td>
<td>35</td>
<td>0.993</td>
<td>0.007</td>
<td>1</td>
</tr>
<tr>
<td>AT1G47490</td>
<td>31</td>
<td>30</td>
<td>17</td>
<td>13</td>
<td>0.79</td>
<td>0.21</td>
<td>0.659</td>
</tr>
<tr>
<td>AT1G48910</td>
<td>27</td>
<td>18</td>
<td>16</td>
<td>27</td>
<td>2.531</td>
<td>-1.531</td>
<td>0.036</td>
</tr>
<tr>
<td>AT1G49290</td>
<td>47</td>
<td>37</td>
<td>12</td>
<td>23</td>
<td>2.435</td>
<td>-1.435</td>
<td>0.04</td>
</tr>
<tr>
<td>AT1G49600</td>
<td>33</td>
<td>9</td>
<td>13</td>
<td>4</td>
<td>1.128</td>
<td>-0.128</td>
<td>1</td>
</tr>
<tr>
<td>AT1G51000</td>
<td>3</td>
<td>13</td>
<td>1</td>
<td>3</td>
<td>0.692</td>
<td>0.308</td>
<td>1</td>
</tr>
<tr>
<td>AT1G54280</td>
<td>121</td>
<td>18</td>
<td>35</td>
<td>23</td>
<td>4.417</td>
<td>-3.417</td>
<td>0.000074</td>
</tr>
<tr>
<td>AT1G55050</td>
<td>71</td>
<td>142</td>
<td>35</td>
<td>49</td>
<td>0.7</td>
<td>0.3</td>
<td>0.187</td>
</tr>
<tr>
<td>AT1G57800</td>
<td>83</td>
<td>102</td>
<td>27</td>
<td>45</td>
<td>1.356</td>
<td>-0.356</td>
<td>0.326</td>
</tr>
<tr>
<td>AT1G57820</td>
<td>96</td>
<td>51</td>
<td>39</td>
<td>38</td>
<td>1.834</td>
<td>-0.834</td>
<td>0.044</td>
</tr>
<tr>
<td>AT1G59670</td>
<td>8</td>
<td>10</td>
<td>2</td>
<td>11</td>
<td>4.4</td>
<td>-0.34</td>
<td>0.129</td>
</tr>
<tr>
<td>AT1G59890</td>
<td>86</td>
<td>119</td>
<td>23</td>
<td>39</td>
<td>1.225</td>
<td>-0.225</td>
<td>0.556</td>
</tr>
<tr>
<td>AT1G60220</td>
<td>48</td>
<td>59</td>
<td>17</td>
<td>12</td>
<td>0.574</td>
<td>0.426</td>
<td>0.213</td>
</tr>
<tr>
<td>AT1G60400</td>
<td>50</td>
<td>72</td>
<td>21</td>
<td>31</td>
<td>1.025</td>
<td>-0.025</td>
<td>1</td>
</tr>
<tr>
<td>AT1G60410</td>
<td>34</td>
<td>42</td>
<td>18</td>
<td>37</td>
<td>1.664</td>
<td>-0.664</td>
<td>0.206</td>
</tr>
<tr>
<td>AT1G60970</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>1.333</td>
<td>0.333</td>
<td>1</td>
</tr>
<tr>
<td>AT1G61090</td>
<td>9</td>
<td>39</td>
<td>5</td>
<td>10</td>
<td>0.462</td>
<td>0.538</td>
<td>0.291</td>
</tr>
<tr>
<td>AT1G61330</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>0.8</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>AT1G62340</td>
<td>102</td>
<td>41</td>
<td>46</td>
<td>39</td>
<td>2.109</td>
<td>-1.109</td>
<td>0.009</td>
</tr>
<tr>
<td>AT1G62660</td>
<td>59</td>
<td>31</td>
<td>59</td>
<td>27</td>
<td>0.871</td>
<td>0.129</td>
<td>0.749</td>
</tr>
<tr>
<td>AT1G63020</td>
<td>122</td>
<td>98</td>
<td>29</td>
<td>39</td>
<td>1.674</td>
<td>0.674</td>
<td>0.072</td>
</tr>
<tr>
<td>AT1G63960</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>0.857</td>
<td>0.143</td>
<td>1</td>
</tr>
<tr>
<td>AT1G67820</td>
<td>66</td>
<td>67</td>
<td>9</td>
<td>26</td>
<td>2.846</td>
<td>-1.846</td>
<td>0.013</td>
</tr>
<tr>
<td>AT1G67830</td>
<td>39</td>
<td>21</td>
<td>20</td>
<td>8</td>
<td>0.743</td>
<td>0.257</td>
<td>0.631</td>
</tr>
<tr>
<td>AT1G69900</td>
<td>33</td>
<td>21</td>
<td>18</td>
<td>17</td>
<td>1.484</td>
<td>-0.484</td>
<td>0.389</td>
</tr>
<tr>
<td>AT1G70560</td>
<td>32</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>6.095</td>
<td>-5.095</td>
<td>0.0032</td>
</tr>
<tr>
<td>AT1G73560</td>
<td>15</td>
<td>12</td>
<td>7</td>
<td>18</td>
<td>3.214</td>
<td>-2.214</td>
<td>0.055</td>
</tr>
<tr>
<td>AT1G76250</td>
<td>59</td>
<td>28</td>
<td>6</td>
<td>11</td>
<td>3.863</td>
<td>-2.863</td>
<td>0.015</td>
</tr>
<tr>
<td>AT1G77000</td>
<td>39</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>2.571</td>
<td>-1.571</td>
<td>0.176</td>
</tr>
<tr>
<td>AT2G17990</td>
<td>22</td>
<td>13</td>
<td>8</td>
<td>10</td>
<td>2.115</td>
<td>-1.115</td>
<td>0.249</td>
</tr>
<tr>
<td>AT2G18880</td>
<td>42</td>
<td>106</td>
<td>7</td>
<td>11</td>
<td>0.623</td>
<td>0.377</td>
<td>0.414</td>
</tr>
<tr>
<td>AT2G19400</td>
<td>25</td>
<td>10</td>
<td>12</td>
<td>11</td>
<td>2.292</td>
<td>-1.292</td>
<td>0.168</td>
</tr>
<tr>
<td>AT2G20160</td>
<td>11</td>
<td>36</td>
<td>3</td>
<td>3</td>
<td>0.306</td>
<td>0.694</td>
<td>0.3229</td>
</tr>
<tr>
<td>AT2G28380</td>
<td>36</td>
<td>20</td>
<td>9</td>
<td>9</td>
<td>1.8</td>
<td>-0.8</td>
<td>0.405</td>
</tr>
<tr>
<td>AT2G31360</td>
<td>14</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2.8</td>
<td>-1.8</td>
<td>1</td>
</tr>
<tr>
<td>AT2G33780</td>
<td>20</td>
<td>41</td>
<td>8</td>
<td>13</td>
<td>0.793</td>
<td>0.207</td>
<td>0.79</td>
</tr>
<tr>
<td>AT2G34880</td>
<td>92</td>
<td>115</td>
<td>12</td>
<td>21</td>
<td>1.4</td>
<td>-0.4</td>
<td>0.451</td>
</tr>
<tr>
<td>AT2G36560</td>
<td>19</td>
<td>53</td>
<td>11</td>
<td>12</td>
<td>0.391</td>
<td>0.609</td>
<td>0.072</td>
</tr>
<tr>
<td>AT3G08040</td>
<td>25</td>
<td>21</td>
<td>13</td>
<td>8</td>
<td>0.733</td>
<td>0.267</td>
<td>0.605</td>
</tr>
<tr>
<td>AT3G10900</td>
<td>22</td>
<td>25</td>
<td>17</td>
<td>29</td>
<td>1.501</td>
<td>-0.501</td>
<td>0.402</td>
</tr>
<tr>
<td>AT3G11160</td>
<td>15</td>
<td>38</td>
<td>1</td>
<td>2</td>
<td>0.789</td>
<td>0.211</td>
<td>1</td>
</tr>
<tr>
<td>AT3G11490</td>
<td>32</td>
<td>13</td>
<td>5</td>
<td>7</td>
<td>3.446</td>
<td>-2.446</td>
<td>0.088</td>
</tr>
<tr>
<td>AT3G14205</td>
<td>56</td>
<td>33</td>
<td>12</td>
<td>9</td>
<td>1.273</td>
<td>-0.273</td>
<td>0.627</td>
</tr>
<tr>
<td>AT3G16320</td>
<td>32</td>
<td>40</td>
<td>10</td>
<td>25</td>
<td>2</td>
<td>-1</td>
<td>0.142</td>
</tr>
<tr>
<td>AT3G17250</td>
<td>40</td>
<td>31</td>
<td>5</td>
<td>14</td>
<td>3.613</td>
<td>-2.613</td>
<td>0.037</td>
</tr>
<tr>
<td>AT3G19160</td>
<td>33</td>
<td>21</td>
<td>11</td>
<td>31</td>
<td>4.429</td>
<td>-3.429</td>
<td>0.0009</td>
</tr>
<tr>
<td>AT3G22180</td>
<td>38</td>
<td>31</td>
<td>20</td>
<td>27</td>
<td>1.655</td>
<td>-0.655</td>
<td>0.256</td>
</tr>
<tr>
<td>AT3G26590</td>
<td>32</td>
<td>14</td>
<td>19</td>
<td>9</td>
<td>1.083</td>
<td>-0.083</td>
<td>1</td>
</tr>
<tr>
<td>AT3G45090</td>
<td>67</td>
<td>41</td>
<td>25</td>
<td>16</td>
<td>1.046</td>
<td>-0.046</td>
<td>1</td>
</tr>
<tr>
<td>AT3G50720</td>
<td>35</td>
<td>45</td>
<td>6</td>
<td>9</td>
<td>1.167</td>
<td>-0.167</td>
<td>1</td>
</tr>
<tr>
<td>AT3G62230</td>
<td>130</td>
<td>164</td>
<td>24</td>
<td>33</td>
<td>1.09</td>
<td>-0.09</td>
<td>0.884</td>
</tr>
<tr>
<td>AT4G00220</td>
<td>13</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>1.486</td>
<td>-0.486</td>
<td>0.694</td>
</tr>
<tr>
<td>AT4G00540</td>
<td>35</td>
<td>55</td>
<td>20</td>
<td>30</td>
<td>0.955</td>
<td>0.045</td>
<td>1</td>
</tr>
<tr>
<td>AT4G01840</td>
<td>37</td>
<td>25</td>
<td>11</td>
<td>10</td>
<td>1.345</td>
<td>-0.345</td>
<td>0.615</td>
</tr>
<tr>
<td>AT4G08590</td>
<td>43</td>
<td>65</td>
<td>13</td>
<td>21</td>
<td>1.069</td>
<td>-0.069</td>
<td>1</td>
</tr>
<tr>
<td>AT4G11940</td>
<td>21</td>
<td>56</td>
<td>17</td>
<td>22</td>
<td>0.485</td>
<td>0.515</td>
<td>0.095</td>
</tr>
<tr>
<td>AT4G13460</td>
<td>62</td>
<td>60</td>
<td>30</td>
<td>29</td>
<td>0.999</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>AT4G15900</td>
<td>32</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>1.322</td>
<td>-0.322</td>
<td>0.746</td>
</tr>
<tr>
<td>AT4G17486</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>0.611</td>
<td>0.389</td>
<td>0.694</td>
</tr>
<tr>
<td>AT4G18650</td>
<td>17</td>
<td>8</td>
<td>17</td>
<td>16</td>
<td>2</td>
<td>-1</td>
<td>0.283</td>
</tr>
<tr>
<td>AT4G20800</td>
<td>45</td>
<td>36</td>
<td>25</td>
<td>25</td>
<td>1.25</td>
<td>-0.25</td>
<td>0.59</td>
</tr>
<tr>
<td>AT4G23110</td>
<td>12</td>
<td>41</td>
<td>5</td>
<td>7</td>
<td>0.41</td>
<td>0.59</td>
<td>0.273</td>
</tr>
<tr>
<td>AT4G26140</td>
<td>41</td>
<td>21</td>
<td>14</td>
<td>13</td>
<td>1.813</td>
<td>-0.813</td>
<td>0.239</td>
</tr>
<tr>
<td>AT4G29650</td>
<td>23</td>
<td>17</td>
<td>6</td>
<td>12</td>
<td>2.706</td>
<td>-1.706</td>
<td>0.155</td>
</tr>
<tr>
<td>AT4G31060</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>1</td>
<td>0.15</td>
<td>0.85</td>
<td>0.165</td>
</tr>
<tr>
<td>AT4G31900</td>
<td>90</td>
<td>195</td>
<td>10</td>
<td>19</td>
<td>0.877</td>
<td>0.123</td>
<td>0.835</td>
</tr>
<tr>
<td>AT4G39955</td>
<td>19</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>0.352</td>
<td>0.648</td>
<td>0.644</td>
</tr>
<tr>
<td>AT5G02630</td>
<td>33</td>
<td>41</td>
<td>16</td>
<td>30</td>
<td>1.509</td>
<td>-0.509</td>
<td>0.342</td>
</tr>
<tr>
<td>AT5G03280</td>
<td>90</td>
<td>31</td>
<td>25</td>
<td>26</td>
<td>3.019</td>
<td>-2.019</td>
<td>0.0023</td>
</tr>
<tr>
<td>AT5G03370</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>14</td>
<td>0.636</td>
<td>0.364</td>
<td>0.704</td>
</tr>
<tr>
<td>AT5G05440</td>
<td>21</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>0.597</td>
</tr>
<tr>
<td>AT5G08360</td>
<td>14</td>
<td>25</td>
<td>3</td>
<td>3</td>
<td>0.56</td>
<td>0.44</td>
<td>0.658</td>
</tr>
<tr>
<td>AT5G10950</td>
<td>55</td>
<td>95</td>
<td>12</td>
<td>26</td>
<td>1.254</td>
<td>-0.254</td>
<td>0.578</td>
</tr>
<tr>
<td>AT5G11460</td>
<td>20</td>
<td>42</td>
<td>4</td>
<td>7</td>
<td>0.833</td>
<td>0.167</td>
<td>1</td>
</tr>
<tr>
<td>AT5G15140</td>
<td>24</td>
<td>12</td>
<td>14</td>
<td>18</td>
<td>2.571</td>
<td>-1.571</td>
<td>0.087</td>
</tr>
<tr>
<td>Gene</td>
<td>Start</td>
<td>End</td>
<td>Length</td>
<td>Max Value</td>
<td>Min Value</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>------</td>
<td>--------</td>
<td>-----------</td>
<td>-----------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>AT5G22920</td>
<td>17</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>2.833</td>
<td>-1.833</td>
<td></td>
</tr>
<tr>
<td>AT5G24460</td>
<td>28</td>
<td>9</td>
<td>21</td>
<td>13</td>
<td>1.926</td>
<td>-0.926</td>
<td></td>
</tr>
<tr>
<td>AT5G28300</td>
<td>32</td>
<td>30</td>
<td>16</td>
<td>33</td>
<td>2.2</td>
<td>-1.2</td>
<td></td>
</tr>
<tr>
<td>AT5G37290</td>
<td>12</td>
<td>6</td>
<td>13</td>
<td>7</td>
<td>1.077</td>
<td>-0.077</td>
<td></td>
</tr>
<tr>
<td>AT5G42670</td>
<td>28</td>
<td>46</td>
<td>13</td>
<td>22</td>
<td>1.03</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>AT5G43780</td>
<td>41</td>
<td>10</td>
<td>59</td>
<td>27</td>
<td>1.876</td>
<td>-0.876</td>
<td></td>
</tr>
<tr>
<td>AT5G46300</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>0.381</td>
<td>0.381</td>
<td></td>
</tr>
<tr>
<td>AT5G50470</td>
<td>28</td>
<td>51</td>
<td>3</td>
<td>10</td>
<td>1.83</td>
<td>-0.83</td>
<td></td>
</tr>
<tr>
<td>AT5G53150</td>
<td>45</td>
<td>76</td>
<td>18</td>
<td>25</td>
<td>0.822</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td>AT5G53870</td>
<td>48</td>
<td>48</td>
<td>15</td>
<td>23</td>
<td>1.533</td>
<td>0.533</td>
<td></td>
</tr>
<tr>
<td>AT1G17770</td>
<td>82</td>
<td>154</td>
<td>30</td>
<td>27</td>
<td>0.479</td>
<td>0.521</td>
<td></td>
</tr>
<tr>
<td>AT1G20730</td>
<td>18</td>
<td>70</td>
<td>2</td>
<td>11</td>
<td>0.155</td>
<td>0.155</td>
<td></td>
</tr>
<tr>
<td>AT2G21450</td>
<td>52</td>
<td>29</td>
<td>24</td>
<td>45</td>
<td>3.592</td>
<td>-1.592</td>
<td></td>
</tr>
<tr>
<td>AT2G40520</td>
<td>34</td>
<td>128</td>
<td>5</td>
<td>12</td>
<td>0.372</td>
<td>0.628</td>
<td></td>
</tr>
<tr>
<td>AT3G19350</td>
<td>11</td>
<td>21</td>
<td>2</td>
<td>14</td>
<td>3.667</td>
<td>-2.667</td>
<td></td>
</tr>
<tr>
<td>AT3G23060</td>
<td>38</td>
<td>82</td>
<td>18</td>
<td>29</td>
<td>0.747</td>
<td>0.253</td>
<td></td>
</tr>
<tr>
<td>AT4G16760</td>
<td>23</td>
<td>5</td>
<td>76</td>
<td>25</td>
<td>1.513</td>
<td>-0.513</td>
<td></td>
</tr>
<tr>
<td>AT4G29570</td>
<td>13</td>
<td>71</td>
<td>8</td>
<td>22</td>
<td>0.504</td>
<td>0.496</td>
<td></td>
</tr>
<tr>
<td>AT4G29640</td>
<td>20</td>
<td>47</td>
<td>16</td>
<td>37</td>
<td>0.984</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>AT5G15140</td>
<td>24</td>
<td>12</td>
<td>14</td>
<td>18</td>
<td>2.571</td>
<td>-1.571</td>
<td></td>
</tr>
<tr>
<td>AT1G52460</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>AT1G59930</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>0.667</td>
<td>-0.333</td>
<td></td>
</tr>
<tr>
<td>AT1G77960</td>
<td>19</td>
<td>30</td>
<td>4</td>
<td>12</td>
<td>1.9</td>
<td>-0.9</td>
<td></td>
</tr>
<tr>
<td>AT3G49770</td>
<td>7</td>
<td>40</td>
<td>12</td>
<td>40</td>
<td>0.592</td>
<td>0.407</td>
<td></td>
</tr>
<tr>
<td>AT3G57250</td>
<td>13</td>
<td>30</td>
<td>2</td>
<td>11</td>
<td>2.38</td>
<td>-1.38</td>
<td></td>
</tr>
<tr>
<td>AT4G18650</td>
<td>17</td>
<td>8</td>
<td>17</td>
<td>15</td>
<td>1.875</td>
<td>-0.875</td>
<td></td>
</tr>
</tbody>
</table>

197
Bibliography


202


