
THE UNIVERSITY
*of* EDINBURGH

# Using Next Generation Sequencing to investigate the generation of diversity in the genus *Begonia*

Katie Emelianova

Doctoral Thesis
Royal Botanic Gardens Edinburgh
University of Edinburgh
2016

## Declaration

I declare that the work contained in this thesis is my own, unless otherwise cited. This thesis has not in whole or in part been previously presented for any degree.

Katie Emelianova
1st September 2016

## Abstract

*Begonia* is one of the most diverse genera on the planet, with a species count approaching 2000 and a distribution across tropics in South America, Africa and South East Asia.
The genus has occupied a vast range of niches; many highly variable growth forms can be found across the distribution, and species exhibit very diverse morphologies, even in closely related species. A recent study has revealed a putative whole genome duplication (WGD) event in the evolutionary history of *Begonia*, which has prompted an interest in investigating the impact gene and genome duplication has had on the diversification of *Begonia*.
To answer questions about phenotypic and ecological diversification in *Begonia*, two species from South America, *B. conchifolia* and *B. plebeja* were chosen as study species based on their close phylogenetic relationship and divergent ecology and phenotype.
RNA-seq data for six tissues from *B. conchifolia* and *B. plebeja* was generated using the Illumina sequencing platform, and normalised relative expression data was obtained by mapping reads to transcripts predicted from the *B. conchifolia* draft genome.
A bioinformatics pipeline was devised to compare expression profiles across 6 different tissues between duplicated gene pairs shared between *B. conchifolia* and *B. plebeja*. Gene duplicate pairs were selected as candidates if they showed divergent expression in one species but not in another. Such duplicate pairs are suggestive of neofunctionalization in one species, providing evidence of a potential basis for phenotypic divergence and diversification between *B. conchifolia* and *B. plebeja*.
Two duplicate pairs were identified as showing such divergent expression patterns as well as being functionally ecologically relevant, Chalcone Synthase and 3-Ketoacyl-CoA synthase, involved in anthocyanin biosynthesis and wax biosynthesis respectively.
Investigation of expression and duplication patterns in both gene families showed the candidate gene families to be strikingly different. While 3-Ketoacyl-CoA synthase showed deeper duplications shared with outgroup taxa, Chalcone Synthase appeared to be expanded very recently, with a burst of duplications specific to the genus.
3-Ketoacyl-CoA synthase showed examples of partitioned expression by tissue for different gene family members, with at least five members of the gene family being highly expressed in one or two tissues only. Chalcone Synthase, however, showed dominance of one basal gene family member. Other Chalcone Synthase members, though expressed at lower levels, showed some evidence of reciprocal silencing in *B. plebeja*, though this pattern was not observed in *B. conchifolia*.
Further investigation of the Chalcone Synthase gene family revealed lineage specific duplication in *B. plebeja*, and more extensive differential duplication patterns were found across other South American *Begonias*. Additionally, signals of positive selection were found in two branches on the Chalcone Synthase phylogeny.

# Lay Summary

The plant genus *Begonia* is very large and diverse, being found across all tropics, with the exception of Australia. Species of *Begonia* grow in a variety of habitats, and are highly varied in appearance and growth form.

Studies in the genus have identified that genome size is very variable across the genus. This variation is in part thought to be due to a high prevalence of genome duplication in the evolutionary history of *Begonia*. Gene and genome duplication is thought to be an important force in the generation of biological complexity and novel traits, however the impact of duplication on the diversity of *Begonia* is not yet clear.

To shed more light on whether duplication of genes allowed *Begonia* to diversify and adapt to many different habitats, this study investigated two species of *Begonia* which are very closely related but are very different in their appearance and habitat. The first, *B. conchifolia*, lives in wet rainforests in the understorey, whereas the second species, *B. plebeja*, lives in more open seasonally dry forests. Leaf morphology is also strikingly different between the two species, *B. conchifolia* having waxy, green and fleshy leaves, and *B. plebeja* has hairier, pigmented leaves.

Next generation sequencing was used to find duplicated genes which had diverged in how they were utilised by one species, but had remained the same in their use by the second species. If duplicated genes showed this pattern, this might mean that the duplication producing those two genes may have helped one species to adapt to a new environment by the evolution of a new or more complex trait.

Two gene duplicates were found which fitted this trend; a gene responsible for producing red pigment, and a gene which is involved with the production of wax coating on plant surfaces. These two genes are considered highly relevant for the study species used here; red pigment acts as protection against a wide range of stresses such as photodamage and insect predators. Waxes covering the surface of plants help prevent drought stress and non-stomatal escape of water.

This study has allowed the identification of genes which may have been important in allowing *B. conchifolia* and *B. plebeja* diversify and adapt, and has provided evidence for gene duplication to be a mechanism that may have contributed to *Begonia* being such a morphologically and species diverse group.

# Acknowledgements

This work could not have been completed without the patience and kindness of my supervisors, Catherine Kidner and Ian Simpson. Their guidance helped me develop ideas within the project, as well as cultivate my interests in the wider scientific sphere.
I'd like to thank students and staff at the Royal Botanic Garden Edinburgh for the wonderful time I spent there, and for all the help along the way. In particular, I'm grateful to Michelle Hart, Laura Forrest and Mark Hughes, who have given me much help along the way.
I'd also like to thank all the staff and students at BioSS, where I had a great deal of support and friendship, and I am very thankful for their kindness and generosity.

Table of contents

# Chapter 1: Introduction

## 1.1 Comparative genomics and utility in non-model organisms

Comparative genomics has become a powerful technique through the rapid development of next generation sequencing (NGS) platforms, and the explosion of genome sequencing in organisms both model and non-model. The widespread availability of genomes, especially from previously unstudied species, has made possible the global comparison of many taxa (Schuster, 2008). We can now ask fundamental questions about the evolutionary histories and strategies employed by species, and how differences between them give rise to the wealth of diversity we see today. Platforms for comparative studies bring together a wide range of functionalities to enable researchers to query both coding and non-coding regions of their genome of interest, and compare against other species. Whole genome sequences for a large number of taxa provide orders of magnitude more coding and non-coding sequence data than previously available, allowing much finer scale and more accurate resolution of phylogenetic relationships.

Comparative genomics can be especially useful in the study of non-model organisms; understanding understudied species can provide greater context for what we know already (McCormack et al. 2013, Tsagkogeorga et al. 2010), as well as being invaluable in conservation efforts and the discovery of new, medically or commercially useful products (Brown et al. 2012, Grivet et al. 2013). Large next generation sequencing (NGS) projects in understudied, non-model organisms can be

problematic due to financial constraints, as well as other limiting factors such as the paucity of genomic resources available. The financial and computational overheads associated with the sequencing of whole genomes is large and often prohibitively so. Other methods are available in the form of reduced representation sequencing (RRS), allowing the capture of a representative portion of the genome (Ma et al. 2017, Blanco-Bercial and Bucklin, 2016). Indeed, many comparative genomics projects do not require a fully sequenced genome, often because they do not require such a large volume of sequence data. RRS methods such as RNA-seq are becoming increasingly used for a wider range of applications (Brereton et al. 2016, Gayral et al. 2013), as are other methods such as RAD-seq (Mastretta-Yanis et al. 2014, Boucher et al. 2016). The former is especially useful as it provides a large volume of sequence data, and expression data associated with it, giving greater context than possible with genome sequencing.

## 1.2 Reduced representation sequencing

Historically, RNA-seq has often relied on the availability of a reference genome to map reads to, and many studies have indeed found that mapping to a genome yields better results than mapping to a reference transcriptome, and better still when annotations are available for the genome (Zhao and Zhang, 2015). Problematic questions are still unresolved, such as whether it is better to map to a reference transcriptome, or the genome of a closely related species, with conflicting reports emerging from the literature (Cahais et al. 2012, Huang et al. 2016). A number of

studies, however, are contributing to a growing evidence base for best practices for assembly and analysis of transcriptome data, providing a better support framework for new studies (Cahais et al. 2012, Gayral et al. 2013, Vijay et al. 2013). Furthermore, the development of new methods such as alignment free mapping (Patro, 2015) is helping to resolve issues associated with transcriptome analyses in non-model organisms, making reference genomes less of a necessity. The phylum Nematoda has benefited greatly from the use of NGS data to resolve phylogenetic relationships. A lack of morphological characteristics and frequent homoplasy has made the systematic resolution of the phylum difficult, and studies using small sequence datasets have struggled to satisfactorily define relationships within the group (van Megen et al. 2009). Complex phylogenetic relationships in the plant kingdom have especially benefitted from RRS, allowing complicated and reticulate relationships to be resolved (Sass et al. 2016, Gardner et al. 2016).

The use of whole genome as well as transcriptome data has enabled much more robust estimation of phylogenies in both subgroups (Desjardins et al. 2013) and in a broader context (Koutsovoulos, 2015). Additionally, the NGS data available for Nematodes has helped to resolve difficult evolutionary histories, such as complex histories of hybridization between species (Lunt et al. 2014).

Rates of evolutionary change can provide insight into the long term evolutionary history of a group. Much has been elucidated about the evolution of avian genomes by comparing different bird lineages, as well as comparing avian genomes with other groups such as mammals. Large scale studies such as Künster et al. (2010) have uncovered patterns of selection and substitution rate variation across a wide range of bird species, while in 2014, Zhang et al. revealed markedly higher selective

constraints across avian genomes compared to mammalian, and provided new insight into genomic regions thought to be instrumental in the functional diversification of bird lineages.

## 1.3 Consequences of gene and genome duplication

The study of polyploidy or whole genome duplication (WGD) has had invaluable contributions from comparative genomics. Polyploidy has shaped the evolutionary history of many groups (Hoegg et al. 2004, Soltis et al. 2012, Jiao et al. 2011), and is widely acknowledged to be a major driving force behind development of complexity and the origins of diversity (Ohno et al. 1968). Understanding how each lineage with a WGD event has responded to the doubling of genetic material in the long term is crucial to understanding how polyploidy and subsequent genome downsizing operate in different selective settings. Comparative genome scale data are therefore necessary to comprehensively assess synteny and gene complements (Koenig and Weigel, 2015). Gene and genome duplication is thought to be a key mechanism driving diversification (Ohno et al. 1968). Generating new genetic material through duplication gives populations the opportunity for increases in complexity and diversity, and processes following duplication shape the fate of gene duplicates. The most likely fate of a duplicate gene is loss through pseudogenization (via rapid accumulation of mutations) or epigenetic silencing (Lynch and Conery, 2000), therefore functional retention of duplicate genes is uncommon. The vast majority of duplicate genes are lost at some point after their birth (Hahn et al. 2007, Demuth and

Hahn, 2009). Immediately after a duplication event, the duplicate pair undergoes a period of relaxed constraint, the result of a lower cost incurred from a mutation due to the degeneracy conferred by the second duplicate. This window can allow the steady accumulation of mutations on one of the duplicates, until a loss of function mutation acts to pseudogenize the gene (Harris and Hofmann, 2015). Though calculating rates of gene loss are difficult because of confounding factors e.g. distinguishing WGD and segmentally derived duplicates, estimates have ranged between 50% and 92% of genes having been lost at some point after duplication (Wagner, 2001). Trajectories of retained genes can vary depending on the circumstances in which they find themselves, and a number of different mechanisms with varying levels of support have been proposed. When duplicated, duplicate genes can undertake new functions in a process called neofunctionalization. A number of examples have been illustrated where, following a decrease in selective constraint, signals of positive selection can be detected in one gene duplicate, suggesting a rapid accumulation of new mutations can cause a duplicate to assume a new function (Pegueroles et al. 2013, Moore and Purugganan, 2003, Lynch and Katju, 2004).

Pre-existing selective constraints can give duplicate genes opportunity to diversify. A gene participating in a number of processes may be under highly purifying selection whilst it's performance in each role may be suboptimal. A duplicate gene can precipitate neofunctionalization via a process known as escape from adaptive conflict (Des Marais and Rausher, 2008), where the redundancy created by the new gene can allow each copy to undergo adaptive change and perform each role more efficiently.

Possible fates of duplicated genes outlined above all provide the opportunity for increases in diversity, at the gene and the species level. Pseudogenization, whilst not

giving rise to new functions, can promote speciation mechanisms via reproductive isolation. Microchromosomal repatterning conferred by duplications can build up isolating barriers sympatrically within species; the divergence of chromosome sequence by differential gene duplication patterns, once different enough, may theoretically form a prezygotic isolating mechanism (Lynch and Conery, 2000). Subfunctionalization too can facilitate increases in complexity, the partitioning of function between duplicates has been associated with neofunctionaliztion also (He and Zhang, 2005). Finally, neofunctionalization, through positive diversifying selection, can give rise to novel, lineage specific adaptations. Such adaptations can contribute to species barriers, promoting speciation and thus increasing species diversity (Monson, 2003). Patterns of duplicate retention and loss through these mechanisms can be compared between species to reveal lineage specific patterns of duplicate evolution. Gene family evolution particular to certain lineages promotes differentiation and diversification, functional biases in retention may indicate aspects of organisms' biology affected by lineage specific duplications (Lu et al. 2012).

Modes of duplication themselves are often governing factors in the loss or retention of duplicates (Blanc and Wolfe, 2004, Aury et al. 2006). Functional biases in retained duplicates from small scale duplications indicate that some functional categories are more likely to be lost than others. Transcription, metabolism and defence are overrepresented in duplicates, suggesting these genes are not dosage sensitive (Gevers et al. 2004). Core housekeeping genes tend to be dosage sensitive because of interaction with other genes and with multigene complexes, and so tend to be longer lived after WGDs, where gene dosage stoichiometry is maintained (Schnable et al. 2011).

While there are examples of phenotype shifts due to divergence in coding sequence of structural genes (Finseth et al. 2015), there is mounting evidence suggesting that expression divergence of duplicate genes may be one of the main routes to phenotypic divergence via duplication. Studies in *Drosophila melanogaster* show that modifications in *cis*-regulatory regions of genes can alter their expression pattern, and thus can yield strikingly different phenotypes. This finding is accompanied by the discovery that the coding region in such genes needn't be altered, and can be functionally equivalent (Li and Noll, 1995). Similar results have been found in mice; a pair of duplicate genes expressed in mouse brain, studying *En-1* and *En-2* (Hanks et al. 1995). Mutations in each duplicate gene were deleterious, however replacement of the coding sequence of one duplicate with another rescued the mutant phenotype. Expression pattern, therefore, was deemed the key difference between the two genes. Studies such as these demonstrate the importance of expression pattern of duplicate genes in creating divergent phenotypes, and studies such as that conducted by Gu et al. (2002) demonstrate the speed and wide extent of expression divergence in duplicated genes.

Much more evidence exists of widespread expression divergence of duplicate genes; surveys of paralogous genes originating from a number of different duplication events in rice has revealed that a large number of genes have become neofunctionalized via divergent expression patterns (Throude et al. 2009, Yim et al. 2009), and further findings supporting these have also been found in other grasses (Yang et al. 2014) and *Arabidopsis* (Blanc and Wolfe, 2004). While it is clear that duplication, especially on a large scale, results in an increase in expression diversity, in many cases a direct link between expression diversification and phenotypic diversification is lacking.

## 1.4 *Begonia*, a model genus for the study of diversity

Figure 1. Representation of *Begonia* morphological diversity

This project has used the large and diverse genus *Begonia* as a model to study patterns of duplications and diversity. With almost 2,000 species (Twyford et al. 2015), it is one of the top ten most diverse Angiosperm genera.

While being a non model species, *Begonia* has attracted much attention due to a large interest in breeding traits of ornamental value (Chen et al. 2015). Wild species of the genus are extremely numerous, as well as being extremely phenotypically diverse. The genus has a pan tropical distribution, being found in Africa (~ 160 species), South and Central America and South East Asia (> 600 species each) (Dewitte et al. 2011), and has a diverse range of plant size, leaf shape, colour and growth form (Brennan et al. 2012).

Divergence of the genus in Africa is proposed to have begun in the Oligocene, based on divergence dates and biogeographic patterns (Goodall-Copestake et al. 2009). Climatic changes are thought to have a role in the early diversification of *Begonia*,

expansion of moist rainforest habitat during the Pliocene and increase in drier conditions later in the Pleistocene may have driven range expansions and contractions, resulting in recurrent patterns of refugia, which is likely to promote speciation (Plana et al. 2004). Climatic oscillations may have driven the evolution of seasonal adaptation in some African species, which have enlarged stems for storing water. Indeed, Goodall-Copestake et al (2010) suggested that if this trait evolved only once, it may have been instrumental in the migration of *Begonia* out of Africa. The phylogenetic relationships of the seasonally adapted African species with American and Asian species of *Begonia* seem to corroborate such a scenario (Goodall-Copestake et al. 2010). Some African species also have fleshy fruits, a trait unusual for the genus, which may also have facilitated long distance migration (Hughes and Hollingsworth, 2008).

Previous research has identified factors such as life history and population structure as contributors to the high species number and morphological and ecological diversity in *Begonia*. Species within the genus have predominantly short-range dispersal, though examples of chance long distance migration do exist (de Wilde et al. 2011). *Begonia* seeds are usually very small and unsuited to animal and wind dispersal, frequently germinating close to the progenitor plant (Twyford et al. 2013). Phylogenetic metrics also point to poor dispersal in *Begonia*, monophyly being frequent in geographically close groups (Hughes and Hollingsworth, 2008). The effects of low dispersal distance are evident in strong population genetic structure seen in the genus, with population substructure being evident over relatively short distances within populations (Hughes

and Hollingsworth, 2008, Hughes et al. 2002). The high frequency of such population structures in the genus suggest that divergence through lack of gene flow and local adaptation over short distances and periods of time can occur with relative ease. Given this, speciation rates would be expected to be elevated in *Begonia*, which may account for the high diversity in the genus. While much population genetic data has been collected in *Begonia*, attention has only recently turned to dissecting genetic and genomic variation and the role it has played in *Begonia*'s diversification.

Previous work suggests that the genus has a highly dynamic genome. Chromosome numbers in *Begonia* are very variable, ranging from 16 to 156 (Neale et al. 2006), indeed the patterns of chromosome numbers within and between taxa have made it difficult to confidently infer a basic chromosome number for the genus (Dewitte et al. 2009). Cytological studies showed significant levels of 2n pollen production, 10 genotypes of a 70 genotype study group showing evidence of unreduced gametes (Dewitte et al. 2010). The same study conducted a progeny analysis, showing that seedlings from 2n pollen were frequent.

Using three *Begonia* species covering a broad taxonomic range in the genus, transcriptomic analysis has revealed a genome duplication early in the evolution of the genus. Brennan et al (2012) used synonymous substitution (Ks) distributions of paralogs from three species of *Begonia* (including *B. conchifolia* and *B. plebeja*) to show evidence for a genome duplication. Paralogous peaks appearing in all three species at similar Ks values suggests that this duplication is shared by the species used for the study. Given the phylogenetic distance between the species, the duplication is likely to have happened near the base of the *Begonia* lineage. Along with this evidence, the occurrence of unreduced gametes, and the frequent viability of offspring obtained from 2n gametes suggests that genome duplication played a significant role

in the evolution of *Begonia*. Likewise, chromosome size varies both between and among groups, Dewitte et al (2009) reported a 12 fold difference between chromosome sizes of *B. dietrichiana* and *B. pearcei*, though inclusion of less than a quarter of the sections of *Begonia* may suggest that even higher levels of variation may be present. A lack of positive correlation between genome size and chromosome number was also found, supporting the evidence for differential chromosome size. Such variable expansion of genomes could be the result of differential accumulation of heterochromatin (Petrov, 2001). Activation of repetitive elements such as transposable elements is a common result of polyploidization (McClintock, 1984), as well as large scale deletion and genome downsizing in the process of genome stabilization (Tate et al. 2009). The studies of chromosome evolution in *Begonia* suggest that these processes have had a major impact on karyotype evolution in the genus, and may have played an important role in the generation of biodiversity in the species.

As well as the genomic and transcriptomic resources presented in this thesis, RBGE has a comprehensive and wide ranging resource base for the investigation of hypotheses relating to Begonia's diversification.

Resources representing species across the genus are available in an extensive living collection, allowing the extraction of DNA for assays looking to ascertain whether identified trends are found in other species in the genus. Additionally, a collection of *B. plebeja* individuals representing a wild species collected by Alex Twyford gives an opportunity for a number of small-scale population genetic analyses in *B. plebeja*, and

could provide valuable insight into the contribution that selection and drift make to any genetic signatures of adaptation found in the studies in this thesis.

Large scale genomic and transcriptomic data originating from a number of studies in *Begonia* is available also; long and short read RNA-seq data is available for the easy access to a wealth of markers for comparative sequence analysis and expression estimation. A draft genome is available for *B. conchifolia*, giving a more complete view of the gene content of a Begonia species, as well as providing access to non-coding regions of the genome.

Finally, other resources for *B. conchifolia* such as genome annotation and a genetic map are available, providing contextual data for following up on transcripts or genomic regions of interest in future studies.

## 1.5 Study species

Two species of *Begonia* were chosen as a model to test hypotheses about duplication and diversity in the genus, *B. conchifolia* and *B. plebeja*. *B. plebeja* is found in seasonally dry forests in Northern Mexico, with a widespread distribution. In contrast, *B. conchifolia* grows in wet tropical forests in Costa Rica, and has a more restricted distribution (Twyford. A, pers. comm.).

This pair of species represent an ideal species pair to investigate hypotheses of duplication and putative adaptive evolution in the genus *Begonia* due to the timing of their divergence.

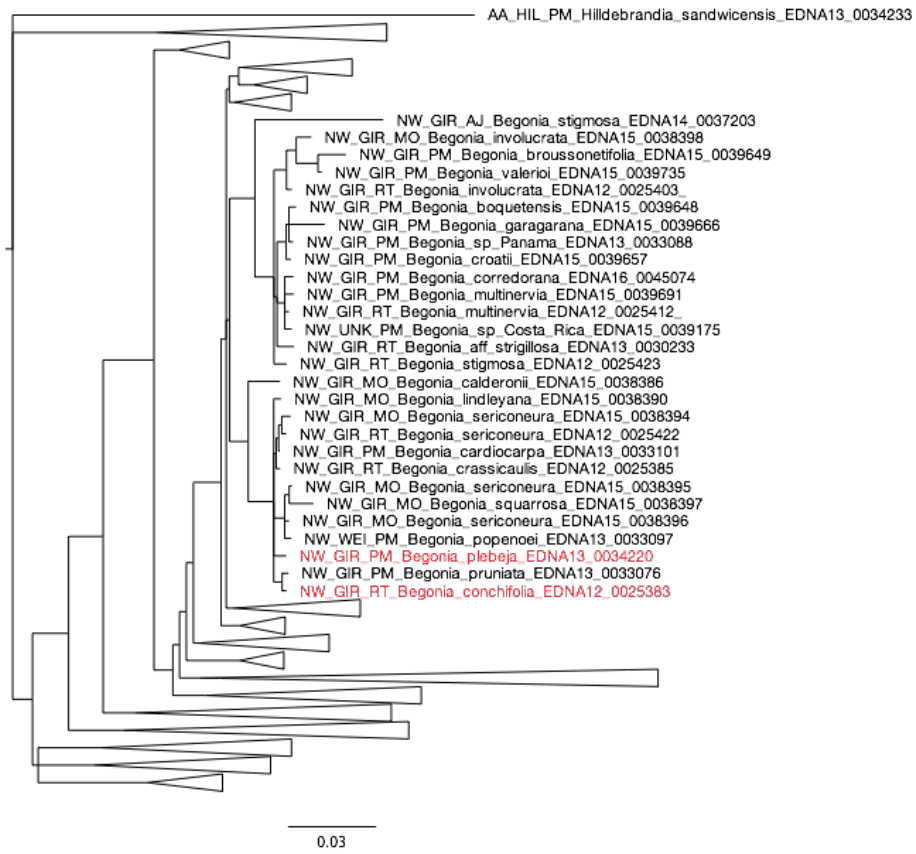



Figure 2. A representative tree showing part of section *Gireoudia*, with the study species *B. conchifolia* and *B. plebeja* shown in red, to illustrate their phylogenetic proximity. Remaining clades have been collapsed to better see relationship of *B. conchifolia* and *B. plebeja*.

On one hand, since speciation the two species have accumulated morphological differences, some of which are likely to be in response to their new and different environments. The presence of such morphological differences suggests that some underlying genetic signature of this adaptation may be detectable. On the other hand, they are very closely related to each other due to the recent divergence time (Figure 2), meaning that any genomic changes underlying adaptation are less likely to be swamped by neutral changes unrelated to the movement away from the ancestral environment into each respective new one.

In addition to some elimination of noise that may have accrued over evolutionary time, the short divergence time makes it highly unlikely that any species-specific whole

genome duplications will have occurred since their divergence. Such an event would have a confounding effect on transcript assembly, copy number estimation and mapping for expression estimation.

Additionally, the historical evidence of polyploidy in wider plant lineages makes *Begonia* a suitable genus to look at polyploidy and its effect on diversity. The closest sequence relative of *Begonia* is cucumber (*C. sativus*), which has no recent history of whole genome duplications (Figure 2). Indeed, the last evidence of polyploidy in the lineage leading to *Begonia* and *Cucumis* is the genome triplication shared by the core Eudicots (diamond in Figure 2). Therefore, the lack of a major genome multiplication since before the divergence of the core Eudicots helps to prevent confounding signal when identifying evidence of gene or genome duplication in *Begonia*.

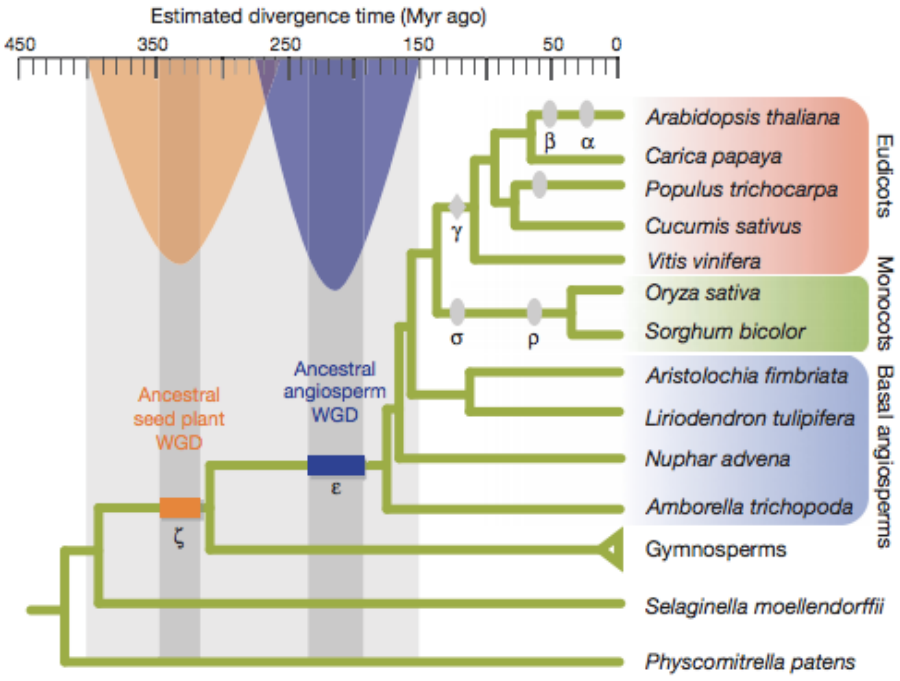



Figure 3. Schematic adapted from Jiao et al. 2011, showing genome duplications (ovals), and triplications (diamond) identified in plant lineages from the literature.

## 1.6 Aims of the thesis

This thesis aims to generate transcriptomic resources for the non-model genus *Begonia*. The availability of a draft genome for *Begonia conchifolia* (see Chapter 2) is an important step in the development of sequence data for *Begonia*, however a wider taxonomic breadth is required to begin to account for the multitude of species the genus contains. Currently, the literature on best practices for transcriptomic studies, especially in non-model organisms, provides a wide ranging set of options and strategies for analysis, providing an excellent opportunity to review the evidence provided by validation studies and appraise results given by suggested methods. Generating good benchmark transcriptomes for *B. conchifolia* and *B. plebeja* has a wider importance for developing the resources needed to study diversity. As one of the most diverse angiosperm genera, both in terms of species richness and phenotypic diversity, *Begonia* represents an excellent model for investigating how diversity is generated, and what factors are important for creating and retaining it, a question important in evolutionary terms, as well as in terms of informing conservation efforts.

To this end, this thesis aims to use transcriptome data from *B. conchifolia* and *B. plebeja*, as well as currently available sequence data from the draft genome of *B. conchifolia*, to help understand the impact gene and genome duplication has had on diversification in *Begonia*. The phylogenetic proximity of the two study species, as well as their shared genome duplication identified in previous studies, provides a system in which the fates of duplicated genes may be tracked, using the relatively short time since speciation as a calibration point. Using a bioinformatics pipeline, gene families will be identified in *B. conchifolia* and *B. plebeja*, using outgroup

species for comparison. Expression and sequence divergence data will be collected across all gene families in an attempt to find gene duplicates which have divergent expression and/or sequence patterns. This strategy will aim to find candidate genes for phenotypic divergence between *B. conchifolia* and *B. plebeja,* which can be studied in more detail.

# Chapter 2: Generation of transcriptomic resources for *Begonia conchifolia* and *Begonia plebeja*

## 2.1 Introduction

### 2.1.1 Transcriptomes and their utility

Transcriptomes are the sequenced collection of total or a subset of RNA within a given sample. The samples can range from whole organisms (Förster et al. 2012) to single cells (Tang et al. 2011). Transcriptome sequencing generates coding sequence of expressed genes in the sample, as well as expression levels of these genes, both of which have an extensive range of applications. They are widely used comparatively, for example examining the effects of different treatments on a sample's gene expression levels (Eierman and Hare, 2015), largely taking over from microarrays for such techniques (Wang et al. 2009). The use of transcriptomes, however, has transcended strictly controlled experiments in model organisms; the reduced representation of a genome they provide makes them excellent for first pass generation of large scale sequence data for any organism (Wei et al. 2011). The utility of transcriptomes in such a capacity is most evident in non-model species with large genomes; retrieval of coding sequence rather than whole genome sequence is often the most cost effective way of obtaining sequence data (Biscotti et al. 2016). Indeed, sequencing of a whole genome is often not necessary for the scientific question a project has, for example where only coding regions of the genome are of interest. The large number of coding sequence transcriptomes produce is attractive to projects which aim to satisfactorily resolve phylogenetic relationships between taxa (McCormack et al. 2013, Rothfels et al. 2013, Hartmann et al. 2012). Groups which have experienced rapid radiations have very shallow

branches towards the tree tips, requiring both very fast, and slower evolving loci, to resolve recent relationships as well as deeper ones, respectively, which can be mined from the range of loci provided by reduced representation sequencing (RRS) technologies such as RNA-seq (Mendoza et al. 2015). Conversely, other groups are difficult to resolve because of a very ancient origin, or due to a complex evolutionary history at the base of the group (Oakley et al. 2013), and here also a wealth of markers has assisted in obtaining the signal necessary to resolve deep relationships. Obtaining a robust phylogeny for a group can further provide context for the markers available from a transcriptome study, and can help to answer complex questions about their evolutionary history (Pease et al. 2016, Sveinsson et al. 2014).

Populations genetics is a field to which transcriptome data has more recently been applied. Transcriptomes have been widely used to obtain sequences to mine for markers such as SNPs and SSRs or to design baits (Hemmer-Hansen et al. 2013, Gugger et al. 2016, Bi et al. 2012, Moghadam et al. 2013). Concerns about using transcriptome data directly for population genetics have been based on problems of hidden paralogy as well as insufficient overlap of data between samples. However, studies have shown that careful appraisal of the limitations of using transcriptome data for population genetics can yield successful results which are comparable to results from analogous studies using genomic data (Gayral et al. 2013, Dlugosch et al. 2013). One very prolific use of transcriptomes, especially in non-model organisms, is candidate gene studies, an invaluable method of leveraging information obtained from model organisms to addressing questions in understudied taxa. The understanding of plant resistance to disease is an important example of having to quickly respond to large scale epidemics in non-model, non-crop species, such as Ash dieback (Harper et al. 2016). The identification of individuals which are resistant to diseases presents an opportunity to identify underlying genetic mechanisms used in resistant individuals. A study by Barakat et al. (2009) used a candidate gene approach to identify genes underlying

resistance to chestnut blight in American and Chinese chestnut, helping to tease apart the genetic basis underlying blight resistance. Comparative transcriptomics is also possible in non-model organisms. The genome wide view of coding sequence provided by transcriptomes can be used to compare taxa with a trait of interest to taxa lacking this trait, uncovering genes putatively associated with the trait (Berens et al. 2015, Biscotti et al. 2016, Sedeek et al. 2013, Yang et al. 2015). Performing such studies is very useful for hypothesis generation, where large scale genomic data is previously unavailable and genome wide patterns are unknown *a priori* (Oppenheim et al. 2014).

### 2.1.2 How transcriptomes can be used to study diversity in *Begonia*

*Begonia* is one of the ten largest genera of angiosperms, comprising almost 2000 species (Moonlight et al. 2015). The reasons behind this specioseness have been investigated using various approaches including population genetics, phylogenetics and cytology (Moonlight et al. 2015, Twyford et al., 2015, Dewitte et al. 2010, Dewitte et al. 2009, Hughes et al. 2003). Observations from a number of studies have shown *Begonia* has poor mechanisms of dispersal, offspring frequently germinating close to the parental plant, corroborating evidence that there is little gene flow between subpopulations of a range (Hughes et al. 2002). Cytological studies indicate *Begonia* has a highly labile genome; great fluctuations in C-value across sections are accompanied by very variable chromosome numbers (ranging from 16 to 156) suggesting genome, chromosome and segmental duplications and rearrangements are frequent (Dewitte et al. 2009). Numerous lines of evidence point to a complex interplay of genetic and environmental factors that contributed to the diversification of *Begonia* and

allowed it to fill many niches in the tropics (Moonlight et al. 2015). The wealth of research on the genus, as well as the highly robust phylogenetic framework available for it, make *Begonia* an excellent model to study how diversity is driven and maintained (Neale, 2006). *Begonia* also represents an understudied aspect of tropical diversity - the plants of the understorey. Maintained plots and air-surveys usually track trees only, the understorey plants can only be surveyed by arduous field work, limiting the information we can gather on them. *Begonia* therefore, with an extensive literature base and a comprehensive research collection, fulfils a role as a model for the study of diversification in the tropical forest understorey.

To achieve a comprehensive resource base for *Begonia*, and to be able to investigate all possible sources and manifestations of the diversity seen in the genus, large scale genomic resources are required. Currently, a draft genome is available for *Begonia conchifolia* (Bombarely and Kidner, Unpublished), as well as a vegetative bud transcriptome for three species *B. conchifolia*, *B. plebeja* and *B. venusta*, sequenced using the Roche 454 platform from single samples (Brennan et al. 2012). This sequence data is a useful resource for investigating *Begonia* at the sequence level, looking for signals of increased rates of evolution and positive selection. However, a dominant force in phenotypic shifts is regulatory change, manifesting as changes in expression patterns of genes, rather than changes in their coding sequence. This mode of phenotype evolution is considered as frequent, if not more so, as evolution via changes at the nucleotide level. Therefore, despite the wealth of sequence resources available for *Begonia*, the absence of a robust set of expression data precludes a comprehensive view of evolutionary processes acting in the genus.

### 2.1.3 Premise of the chapter

This chapter will describe the generation of expression data, which will be used in the investigation of gene expression divergence in a pair of closely related though ecologically distinct *Begonia* species. Two species of *Begonia* have been selected for use in this study. The first, *B. conchifolia*, is an understorey herb from wet forests in Central America, which prefers shaded and moist habitats (Figure 1). Genetic resources for *B. conchifolia* include a draft genome (Bombarely and Kidner, Unpublished) and a vegetative bud transcriptome sequenced using the Roche 454 platform (Brennan et al. 2012), as well as a genetic map. The second species is *B. plebeja*, from the seasonally dry forests of Mexico and central America (Figure 2). *B. plebeja* has a genetic map and a Roche 454 vegetative bud transcriptome (Brennan et al. 2012). Despite the divergent morphology and habitat preference between the two species, they diverged very recently - between 3 and 5 MYA (Moonlight et al. 2015). The disparity between the short divergence time and high environmental divergence is representative of the high species and subsequent morphological diversification found across *Begonia*, and readily available living collections of the species makes this species pair an excellent study system. This project will investigate expression across different tissues in *B. conchifolia* and *B. plebeja*, using Illumina short read technology to examine how the species allocate expression across the whole plant, using biological replicates to enable robust statistical analysis of relative expression. 454 transcriptomes used to find markers for the production of the genetic map were made from vegetative buds. This was to increase the chances of finding markers associated with the differences in vegetative form between the two species. For this more detailed analysis of the differences between the species, six tissues were used: female flower, male flower, mature leaf, mature petiole, vegetative bud and roots.

Figure 1. *Begonia conchifolia*

Figure 2. *Begonia plebeja*

Although the species differ mostly in their vegetative forms, there are some differences in flowers, principally in size, which have been linked to QTLs (Twyford et al. 2015). Comparisons between female and male flowers may also identify genes involved in development of single sex flowers typical of *Begonia*. The mature leaf was chosen to examine differences in the gene expression which could be linked to adaptation to the different light and water resources of the environments the two species grow in. The petiole samples provide a non-photosynthetic tissue control for this. The vegetative bud samples may identify genes regulating the production of different leaf forms and different growth patterns. Very little work has been done on *Begonia* roots, therefore it is not yet known to what extent they will vary between species. Although the roots used for this study are grown in the same environment (pot grown in bark-based compost), they may show variation in pathways for mineral and water uptake important to the plants.

## 2.2 Methods

### 2.2.1 Genome sequencing

Genome sequencing was carried out outwith the work of this thesis; DNA extraction was performed by Catherine Kidner at Royal Botanic Gardens Edinburgh, and sequencing and bioinformatics analyses were carried out by Aureliano Bombarely at Virginia Tech.

The draft genome of *B. conchifolia* was sequenced and annotated for use as a reference in *Begonia* evolutionary studies (Bombarely and Kidner, Unpublished).

Two different sequencing platforms were used to produce reads for the genome, Illumina, and Pacific Biosystems (PacBio). PacBio reads were generated using 4ug of genomic DNA from *B. conchifolia* (Accession Number: 20042082) at Genome Quebec at MacGill University; a large insert library (SMRTbell) was prepared following manufacturers instructions, and sequenced according to facility specifications. The PacBio reads produced here were used in the final assembly with an additional 38.96 Gb of paired end Illumina reads. Illumina reads were generated by Edinburgh Genomics; an Illumina TruSeq DNA kit was used to prepare libraries, which were sequenced on a HiSeq 2500 Illumina System. Illumina reads were corrected using Musket (Liu et al. 2013) and duplicates were filtered using Prinseq (Schmieder et al. 2011). PacBio reads were corrected with the help of Illumina reads using LoRDEC (Salmela and Rivals, 2014).

### 2.2.2 Genome assembly

All computational analyses were done using a 64 thread Red Barn server (Ubuntu 14.04) with 128 GB RAM and 3Tb Hard Drive. A hybrid approach was used to assemble data from the two sequencing platforms. Two base assemblies were generated; Illumina reads were

assembled using SOAPdenovo2 (Luo et al. 2012) using an optimised kmer length ranging from 31 to 95, and PacBio data was assembled using Sprai using default parameters. Each assembly was improved using sequence data from the other technology; The Illumina assembly was improved using PacBio reads, using SSPACE-Long (Boetzer and Pirovano, 2014) for rescaffolding and PBJelly (English et al. 2012) for gap filling, and The PacBio assembly was improved using Illumina reads, using SSPACE (Boetzer et al. 2011) for rescaffolding and GapCloser (Luo et al. 2012) for gap filling. Final assemblies were filtered for plastid sequences and sequences below 1KB in length. The two resulting assemblies were then compared using assembly statistics such as total size, longest sequence, N90/L90, N50/L50 and gap size. Based on these statistics, the PacBio assembly with Illumina read improvement was selected as the best assembly.

### 2.2.3 Genome annotation

The *B. conchifolia* genome was structurally annotated using MAKER-P (Cantarel et al. 2008) using default parameters. An initial transcriptome assembly was generated using 16 RNA-seq libraries (see below) using Trinity (Grabherr et al. 2011). Augustus (Stanke et al. 2005) and SNAP (Johnson et al. 2008) were used within MAKER-P, and both were trained with the draft transcriptome assembly using default parameters. Repetitive sequences were extracted from the genome using RepeatModeler (Smit et al. 2014).

MAKER-P was then used to annotate the genome using extracted repeat sequences, the draft *de novo* transcriptome assembly and trained *ab initio* gene predictor output files.

Identifiers in the annotation files were modified with a Perl script following the format: Five letters species prefix + Assembly version + Three letters sequence type (Scf or Ctg) +

Sequence number + Single letter annotation type (e.g. “g” for gene) + gene number based in the position in the sequence. Functional annotation was performed by sequence homology search using BLASTP, with a minimum E value of $1e^{-10}$, with GenBank, TAIR and SwissProt protein datasets. Additionally, InterProScan was used to annotate protein domains, extending the annotation to Gene Ontology terms associated with these protein domains. Functional descriptions were processed using AHRD giving a weight of 100, 50 and 30 to SwissProt, TAIR and GenBank annotation respectively.

2.2.4 RNA extraction

The tissues chosen for study, mature leaf, mature petiole, vegetative bud, female flower, male flower and root, were harvested between 9am and 10am between January and May 2015 from *B. conchifolia* (Accession Number: 20042082) and *B. plebeja* (Accession Number: 20051406). Leaves were the first fully expanded leaf on the axis, petioles were from these leaves, flowers were between tepals just opening and tepals fully expanded, roots were young white roots within 5-10 cm of the apex. Tissue was frozen in liquid nitrogen immediately upon collection. Plant material was ground finely with liquid nitrogen using a mortar and pestle, adding between 100 and 200 mg of frozen ground tissue to 500ul of Plant RNA extraction reagent (Invitrogen). The tube was vortexed vigorously and incubated at room temperature for 5 minutes horizontally to increase contact area with air. After incubation, the samples were spun for 2 minutes at $4^{o}$C and 12,000 x g. Spun samples were transferred to new tubes with 100ul 5M NaCl, being careful not to disturb the pellet, and 300ul chloroform added. Samples were vortexed vigorously and spun for 10 minutes at $4^{o}$C and 12,000 x g. The liquid phase from samples was transferred to new tubes with 500ul acid phenol chloroform (Fisher, pH 4), vortexed vigorously and spun for 10 minutes at $4^{o}$C and 12,000 x

g. The liquid phase from samples was transferred to an equal amount of isopropanol, vortexed vigorously and incubated at -18$^{o}$C for 2 hours. After incubation, samples were spun for 10 minutes at 4$^{o}$C and 12,000 x g, and supernatant discarded. The resulting pellet was washed with 1ml 75% ethanol, spun for 1 minute at 4$^{o}$C and 12,000 x g, and ethanol discarded. The pellet was resuspended in 12 - 15ul of DEPC treated dH2O, pipetting up and down several times to ensure the pellet was dissolved. After quantification, samples were stored at -80$^{o}$C. RNA was quantitated using Qubit (Thermo Fisher) according to manufacturers instructions, using RNA Broad Range reagents and settings. Sample purity (260/280 ratio) was estimated using a NanoVue Spectrophotometer according to manufacturers instructions.

### 2.2.5 RNA Sequencing

Library preparation and sequencing, carried out by Edinburgh Genomics, consisted of preparation of TruSeq mRNA-seq libraries, and generation of 240 million 150 base pair paired-end reads on one lane of a HiSeq rapid v1 machine. Raw reads are stored in the European Nucleotide Archive.

### 2.2.6 Quality control

Raw reads were trimmed using Trimmomatic 0.36 (Bolger et al. 2014) using a 4 base sliding window and a minimum mean quality of 15. Leading and trailing bases lower than quality score 3 were trimmed. Trimmed FASTQ files were quality checked using FastqC 0.11.5 (Patel and Jain, 2012).

### 2.2.7 Transcriptome Assembly

The *de novo* assembly software Trinity 2.4.0 (Grabherr et al. 2011) was used for transcriptome assembly. The number of reads from each tissue together constituted a large amount of data, and to find the optimal size of data input for Trinity, three different assemblies were made. The first was obtained by running Trinity with default parameters on all reads from all tissues and replicates. The two other assemblies involve assembly of tissues and assembly of replicates independently using Trinity and default parameters. After this, a number of steps are taken to remove redundant sequences from the concatenated file of tissue or replicate assemblies. First, the longest sequence from each group of isoforms is taken, discarding the remaining sequence, and the retained sequence is filtered by length, discarding all transcripts which are shorter than 300 bp. A custom python script (available at https://github.com/katieemelianova/Pipelines/blob/master/pipelineGenericNoClusterNoSelectionBegoniaDistOnly.py) was used to merge clusters of sequences which shared over 99% identity across at least 200 bp. Finally, the remaining sequences were filtered for redundant sequences using the tr2aacds pipeline in the EvidentialGene package (Gilbert, 2013).
The three assemblies were compared for completeness and fragmentation by looking at assembly metrics such as total number of sequences, longest sequence, N50 and N90, and the best assembly was selected for use as a reference transcriptome. BUSCO v2 (Simão et al. 2015) was used to check transcriptome completeness and extent of fragmentation, using the transcriptome setting against the Embryophyta odb9 lineage set of single copy orthologs. Transcriptomes assembled representing each tissue from *B. conchifolia* and *B. plebeja* were also kept for further analyses.

### 2.2.8 Annotation

Transcripts were annotated by transferring annotations of predicted protein sequences from the *B. conchifolia* draft genome. For *B. conchifolia,* all reference transcriptome sequences were queried against the *B. conchifolia* predicted transcripts using BLASTN 2.2.3 using default parameters and no E value cutoff. Hits which had 99.5% identity across at least 300 bp were used to identify each reference transcript's genomic counterpart, using this % identity cutoff to account for sequencing errors. A similar strategy was used to do the same in *B. plebeja*, with a 98% identity cutoff being applied. Annotations were then transferred to reference transcripts which were successfully mapped to predicted transcripts from the *B. conchifolia* genome.

### 2.2.9 Read mapping and transcript quantification

Trimmed and quality checked reads from each tissue and replicate were mapped to the selected reference transcriptome (see Results) and transcripts were quantified using Salmon 0.7.2 (Patro et al. 2015), an alignment-free transcript quantification tool giving rapid and accurate transcript abundances. Default parameters were used and the raw read counts were used for further expression analyses (see 2.3.0).

### 2.3.0 Differential gene expression

EdgeR (Robinson et al. 2010) was used to identify differentially between pairwise comparisons of all tissues. Relative transcript abundance was calculated using the TMM method to normalise across samples. Differentially expressed genes were identified using the GLM likelihood ratio test implemented in EdgeR 3.16.5, performing each test pairwise between each tissue/species combination by specifying every comparison as a contrast in each GLM LRT test. Genes were identified as significantly differentially expressed if they were significant at alpha = 0.05, adjusting the P-value using the FDR method.

## 2.3 Results

### 2.3.1 Transcriptome Assembly

| Assembly | N | Min | Max | Bases | Mean bp | < 200 bp | > 1K bp | ORF | n90 | n50 |
|---|---|---|---|---|---|---|---|---|---|---|
| **CON all** | 100703 | 224 | 16161 | 67065262 | 665.97084 | 0 | 16624 | 23552 | 282 | 1004 |
| **CON individual** | 37533 | 301 | 15544 | 53111540 | 1415.06248 | 0 | 20525 | 25340 | 684 | 1897 |
| **CON tissue** | 20562 | 301 | 15561 | 35114785 | 1707.75143 | 0 | 14052 | 17704 | 910 | 2159 |
| **PLE all** | 121450 | 224 | 15565 | 73405529 | 604.40946 | 0 | 16688 | 24723 | 269 | 821 |
| **PLE individual** | 68232 | 301 | 11720 | 89395841 | 1310.17471 | 0 | 36003 | 49987 | 654 | 1718 |
| **PLE tissue** | 19798 | 301 | 15517 | 30477453 | 1539.4208 | 0 | 12101 | 16434 | 785 | 2014 |

Table 1. Assembly metrics for all three assemblies in *B. conchifolia* and *B. plebeja*

| | % Complete | % Complete Single Copy | % Complete Duplicated | % Fragmented | % Missing |
|---|---|---|---|---|---|
| *B. c* female flower | 83.5 | 60.4 | 23.1 | 4.8 | 11.7 |
| *B. c* leaf | 80.5 | 62.6 | 17.9 | 6.0 | 13.5 |
| *B. c* male flower | 84.5 | 53.5 | 31.0 | 5.1 | 10.4 |
| *B. c* petiole | 83.4 | 56.9 | 26.5 | 5.3 | 11.3 |
| *B. c* root | 84.7 | 52.3 | 32.4 | 5.1 | 10.2 |
| *B. c* veg. bud | 87.3 | 62.5 | 24.8 | 5.0 | 7.7 |
| *B. p* female flower | 81.0 | 49.5 | 31.5 | 6.6 | 12.4 |
| *B. p* leaf | 83.1 | 48.3 | 34.8 | 6.0 | 10.9 |
| *B. p* male flower | 76.1 | 49.0 | 27.1 | 7.9 | 16.0 |
| *B. p* petiole | 84.0 | 51.6 | 32.4 | 5.4 | 10.6 |
| *B. p* root | 70.2 | 47.8 | 22.4 | 10.7 | 19.1 |
| *B. p* veg. bud | 79.1 | 54.0 | 25.1 | 7.4 | 13.5 |

Table 2. Percentage of single copy ortholologs from OrthoDB in five categories reflecting *Begonia* transcript completeness obtained from a BUSCO analysis. Total number of single copy orthologs used is 1440.

Three assemblies were made for *B. conchifolia* and *B. plebeja* (Table 1): Trinity was first used to assemble all reads from all tissues and replicates, hereafter referred to as the 'all' assembly. The second assembly strategy, referred to as the 'replicates' assembly also used Trinity, making assemblies from each individual replicate from each tissue, resulting in 18 assemblies per species. These were then merged and duplicate transcripts were removed using the tra2aacds pipeline in the EvidentialGene suite of programs, designed to remove duplicate sequence from merged or hybrid assemblies. The third assembly strategy, the 'tissue' assembly, used the same method as the second strategy, instead using reads from all replicates for each tissue for each assembly, producing 6 assemblies for each species which were then also merged using the tr2aacds pipeline. The first strategy proved to be the worst, producing a highly fragmented assembly with over 100,000 sequences. A markedly lower N50 for the assemblies using all reads compared to the other two methods suggests that the assembly is composed of much shorter sequences than the tissue and replicate assemblies. While this metric is used mainly for assessing the contig lengths of genomes, the much lower N50 illustrates the disparity between the methods.

Analysis of the tissue transcriptomes using BUSCO revealed around 80% of single copy orthologs from the Embryophyte lineage were complete in each of the tissues sequenced in both *B. conchifolia* and *B. plebeja*, suggesting that most genes are captured successfully. Any missing genes can be attributed to either too low coverage resulting in the missing of the genes from the assembly, or the tissue specific expression of the gene, this being reflected in a roughly 10% rate of single copy orthologs being missing from each tissue. Technical reasons for missing or fragmented regions of genes are likely to underlie at least some of the

missing or fragmented genes; singleton reads may be discarded or closely related paralogs or regions of paralogs could be collapsed.

The number of bases assembled in the ‘all’ assembly is similar to the ‘replicates’ assembly, though both are around 2 fold greater than the ‘tissues’ assembly. The highly increased number of bases in the ‘all’ and ‘replicates’ assembly indicates that a large number of misassemblies may be present within them, and that much sequence has not been aligned and collapsed sufficiently (Table 1). Meanwhile, the longest sequence for all assemblies is relatively similar, however this is not a very revealing metric and may reflect longer and easier to assemble transcripts. A reduction in the number of sequences with an ORF as predicted by Transrate (Smith-Unna et al. 2016) in the ‘tissues’ assembly is suggestive of the ‘replicates’ and ‘all’ assembly having a much larger number of sequences, and an artificially inflated number of ORFs being found in these due to internal start and stop codons being misinterpreted as full ORFs. BLASTX was used to query *A. thaliana* primary transcript peptide sequences using *B. conchifolia* and *B. plebeja* reference transcript sequences from each assembly method, retaining only hits of over 70% identity across at least 300 base pairs. Plotting the lengths of *Begonia* query length against the length of the corresponding *A. thaliana* target length supported the 'tissue' assembly being the best of the three (Figure 3). Fewer hits representing a long *A. thaliana* sequence matching a much shorter *Begonia* sequence were seen in the 'tissue' assembly, however *B. plebeja* appeared to have more sequences with a visibly greater proportion being shorter than their counterparts in *A. thaliana*. Given the transcriptomes were treated equally, this cannot be attributed to methodological biases introduced in bioinformatics analysis. Estimating the density of *B. conchifolia* and *B. plebeja* sequence lengths (figure 4) showed a very high density of sequences with very short lengths in 'all' assemblies from both species. A smaller disparity

was seen between 'tissue' and 'replicate' assemblies, but both species showed clearly that a skew towards longer sequences was demonstrated by the 'tissue' assembly.

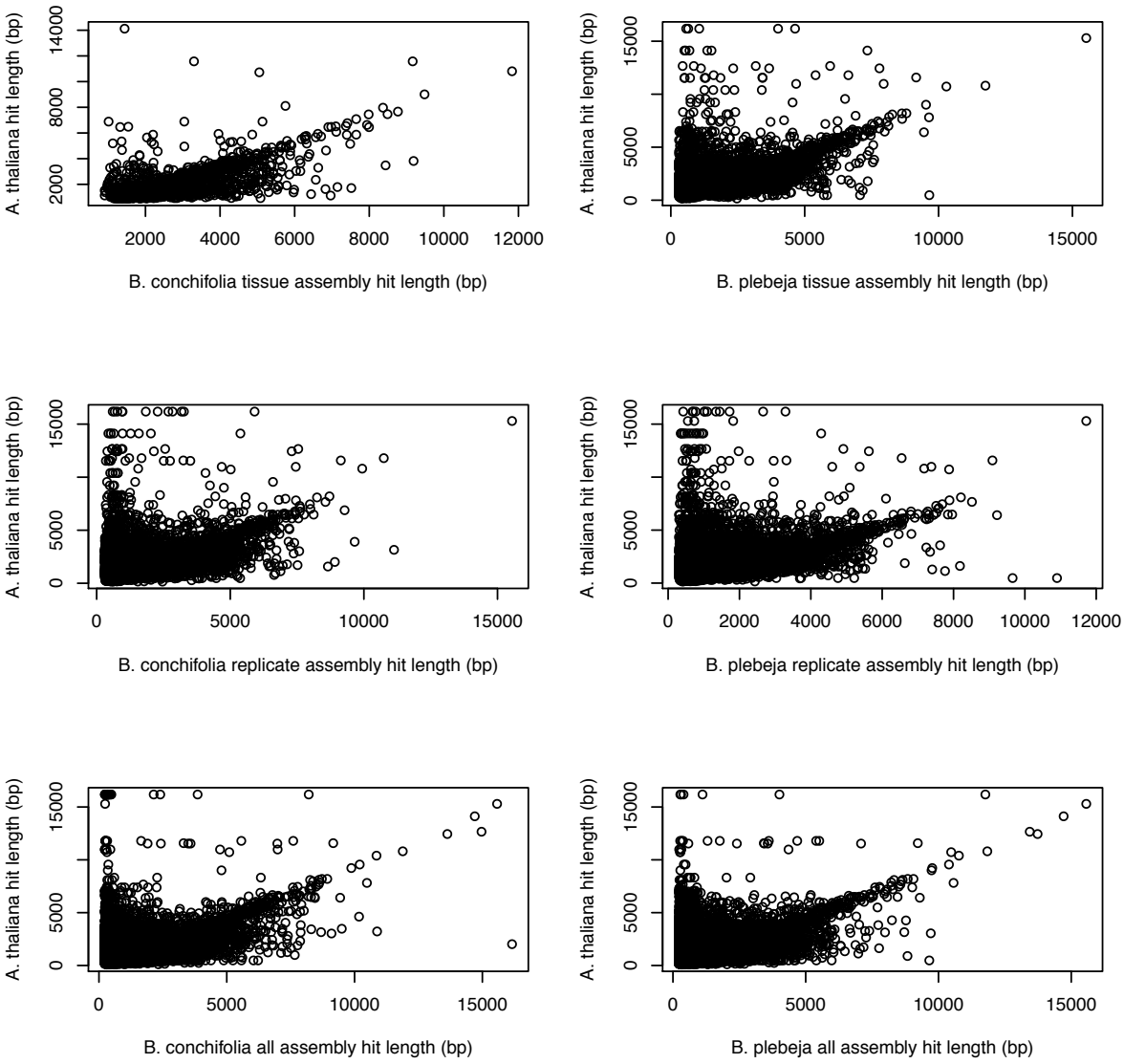



Figure 3. Hit length of *B. conchifolia* and *B. plebeja* hits to *A. thaliana* proteins for All, Individual and replicate assemblies. BLASTX was used to search the TAIR10 protein database at an E value threshold of $1e^{-40}$

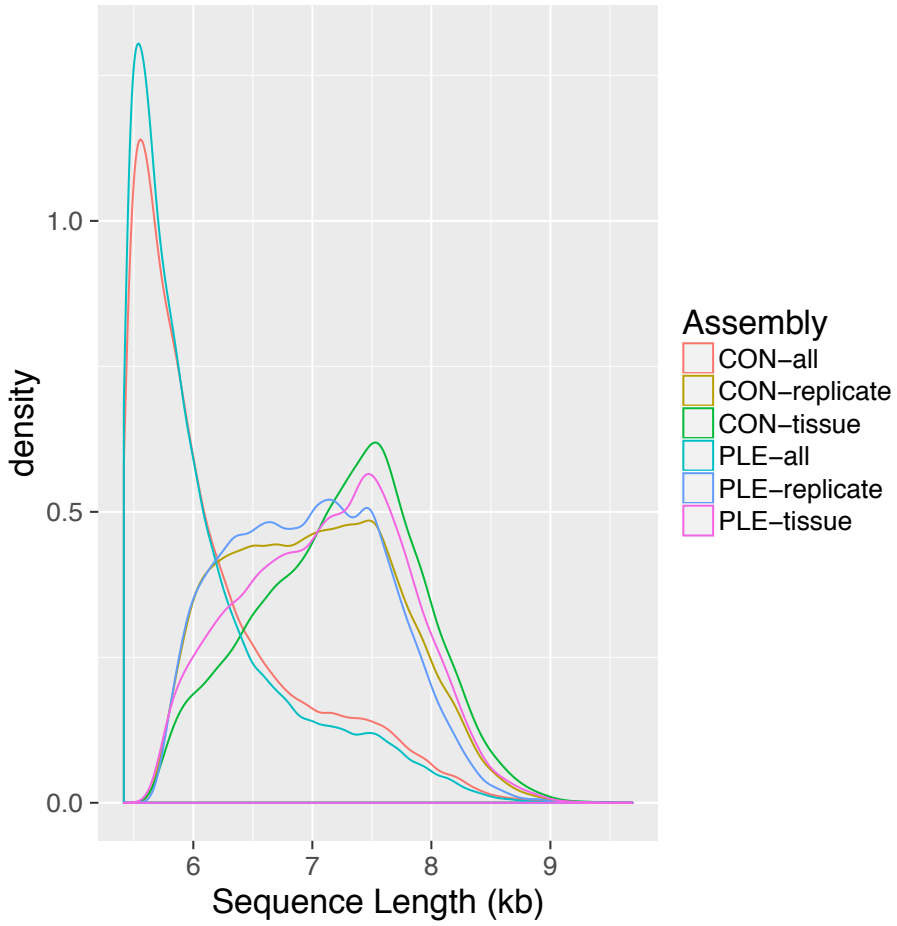



Figure 4. Density distribution of all three assembly methods in *B. conchifolia* and *B. plebeja*

2.3.2 Tissue expression patterns

| Tissue | *B. conchifolia* | | | *B. plebeja* | | |
|---|---|---|---|---|---|---|
| | **Rep 1** | **Rep 2** | **Rep 3** | **Rep 1** | **Rep 2** | **Rep 3** |
| **Female Flower** | 1904 | 1852 | 1876 | 1931 | 1893 | 1865 |
| **Leaf** | 1516 | 1426 | 1419 | 1336 | 1336 | 1482 |
| **Male Flower** | 1454 | 1560 | 1541 | 1664 | 1755 | 1742 |
| **Petiole** | 1851 | 1750 | 1849 | 1759 | 1744 | N/A |
| **Root** | 1796 | N/A | 1792 | 1866 | 1872 | 1811 |
| **Vegetative Bud** | 1811 | 1762 | 1832 | 1853 | 1900 | 1867 |

Table 3. Number of transcripts identified as being expressed in each species for each tissue and replicate. Expression of a feature is defined as having greater than 100 counts per million. N/A denotes tissues which were eliminated from analysis due to technical difficulties

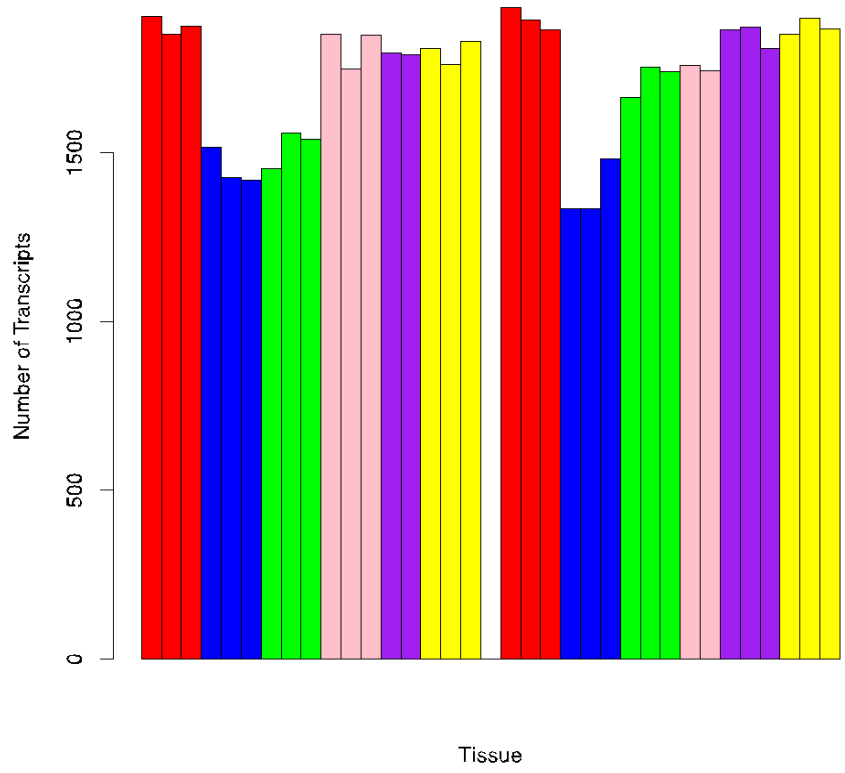



Figure 5. Barplot of number of transcripts identified as being expressed in each tissue. Bar cluster on left corresponds to *B. conchifolia*, cluster on right to *B. plebeja*. Colours correspond to: red = female flower, blue = mature leaf, green = male flower, pink = petiole, purple = root, yellow = vegetative bud.

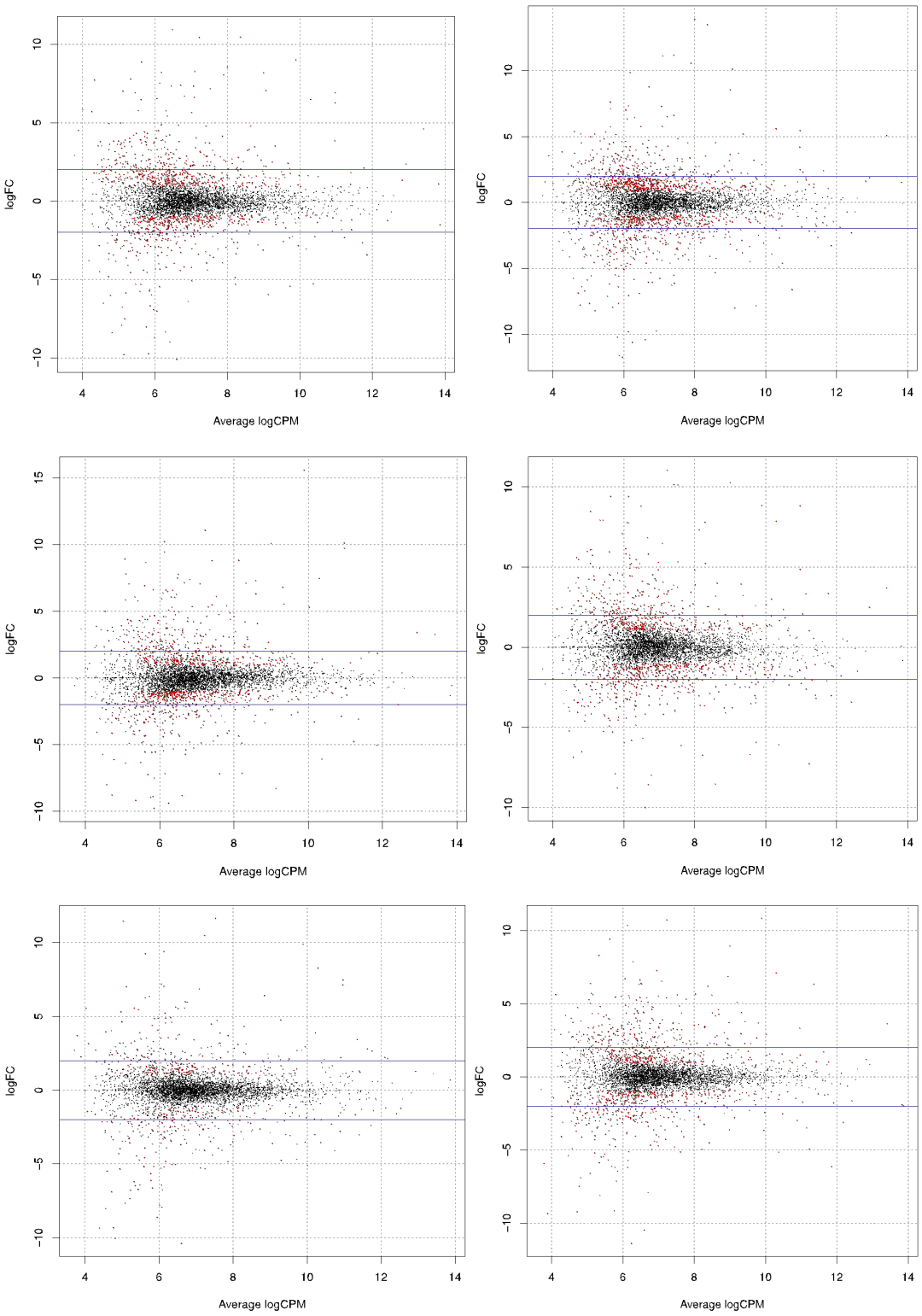



Figure 6. MA plots of differentially expressed genes between *B. conchifolia* and *B. plebeja* for each tissue. Top row left = female flower, top row right = male flower, mid row left = leaf, mid row right = petiole, bottom row left = root, bottom row right = vegetative bud. Points coloured in red indicate significantly differentially expressed sequences (alpha = 0.05 with FDR correction), blue lines mark 2.5 fold change.

Number of genes expressed in each tissue were comparable between *B. conchifolia* and *B. plebeja* in petiole, root and vegetative bud (figure 5). Leaf in both species had the least number of expressed genes, while female flower had the most. Petiole, female flower and vegetative bud had the next most number of expressed transcripts in both species, expressing a largely similar number of genes on average between these tissues as well as between *B. conchifolia* and *B. plebeja.* Interestingly, the number of genes expressed in female and male flowers showed a relatively large discrepancy in both species, male flower being more similar in number of expressed genes to leaf than to female flower in both species, though male flower had a slightly elevated number of transcripts in *B. plebeja* compared to *B. conchifolia*. MA plots (figure 6), showing the counts per million (CPM) of each locus plotted against the fold change (logFC) between *B. conchifolia* and *B. plebeja*, showed for each tissue how many genes were expressed at least 2 fold higher in *B. conchifolia* compared to *B. plebeja* (points above the top blue line) and at least 2 fold lower in the same two species (points below bottom blue line). Loci which are significantly differentially expressed at alpha = 0.05 (corrected for multiple testing using the FDR method) are shown in red, and those not significantly differentially expressed in black. A similar number of genes were both upregulated and downregulated in *B. conchifolia* compared to *B. plebeja* overall (figure 7, appendix 3). Roots had the least number of genes that had at least a 2 fold difference, 186 were upregulated in *B. conchifolia* while 153 were upregulated in *B. plebeja*. The highest differences were seen in male flower; 298 genes were upregulated in *B. conchifolia* compared to 258 in *B. plebeja*. Vegetative bud, petioles and leaves had similar numbers of loci with both upregulated and downregulated genes, ranging from 210 to 260. Female flower, while having a similar number of loci upregulated in *B. plebeja*, had a much lower number of loci upregulated in *B. conchifolia* at 178.

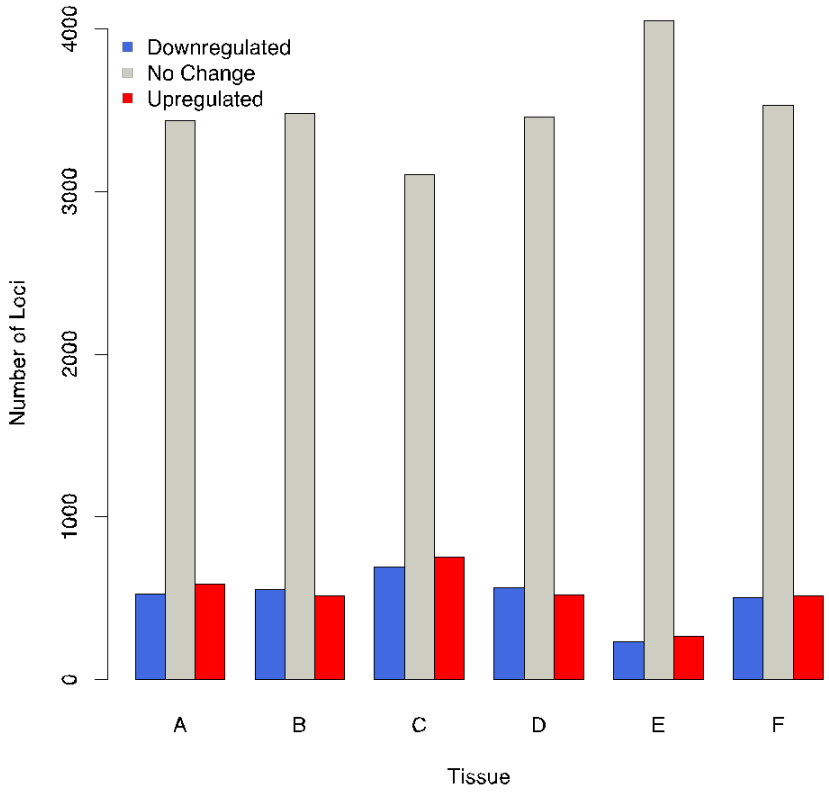



Figure 7. Barplot of up- and down-regulated loci between tissues of *B. conchifolia* and *B. plebeja*. Upregulated denotes upregulation of *B. conchifolia* relative to *B. plebeja* and downregulation denotes the reverse. Tissue codes are: A = female flower, B = leaf, C = male flower, D = petiole, E = root, F = vegetative bud

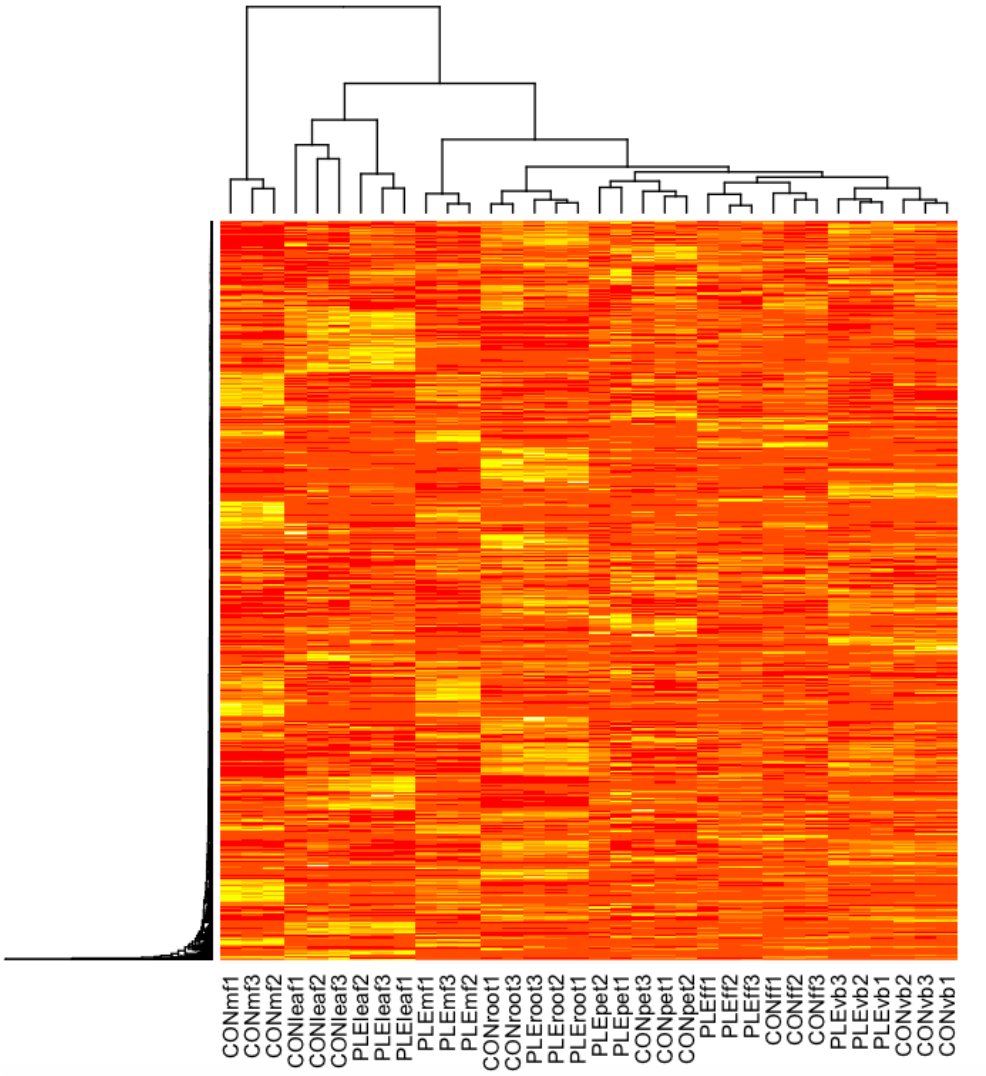



Figure 8. Heatmap of expression across tissues and replicates in *B. conchifolia* and *B. plebeja*. Difference between features is calculated using Euclidian distance, relationships of samples is calculated using hierarchical clustering. Normalised expression values were used as input (see section 2.3.0)

For significantly differentially expressed genes (fold change > 5), Mercator (Lohse et al. 2014) was used to visualise the distribution of functional categories in each comparison. For genes differentially expressed in each tissue, pie charts giving proportions of functional terms in each comparison are provided in appendices 16 to 21.

Using hierarchical clustering, a heatmap was created using the normalised expression values obtained by mapping of tissue and replicate reads to the species specific reference transcriptomes and normalising using EdgeR (see section 2.3.0). The heatmap allows visualisation of relationships between tissues in terms of expression, as well as the amount of heterogeneity within each tissue between replicates (figure 8).

Petiole, female flower and vegetative bud clustered well together both within species and between, and were more similar to each other in expression. Interestingly, vegetative bud was most similar to female flowers, rather than male flowers as would be expected. Roots were also similar to these three tissues, though more distantly, but clustered together between and within tissues well. Leaves were much less similar to the former four tissues; while clustering together, the leaf tissue expression fell well outside the group formed by vegetative bud, root, female flower and petiole.

Unexpectedly, male flowers clustered together within species, but not between species; *B. conchifolia* male flower was found to be most distantly related to all other tissues, whereas *B. plebeja* male flower expression clustered next to the group containing petiole, root, female flower and vegetative bud.

### 2.3.3 Shared expressed genes

A similar number of sequences are specific to either *B. conchifolia* or *B. plebeja* (~5000 in each), a much smaller number than *C. sativus* (17404) and *A. thaliana* (24969), which likely reflects the incompleteness of the *Begonia* transcriptomes and the quality of annotation and high coverage of *C. sativus* and *A. thaliana*, particularly the latter (figure 9). While the *Begonia* transcriptome sequence is not complete, some useful comparisons are able to be made. 23,133 sequences are shared by the two *Begonia* species, though some sequences may have matched several times or have multiple matches due to gene family expansion. The loci which appear to be unique to *B. conchifolia* and *B. plebeja* respectively may represent novel or highly divergent genes, though the possibility that these are genes that are shared by the *Begonia* species but are missing from the transcriptome assembly due to low expression can not be discounted. A much larger set of genes is shared by *Begonia* and *C. sativus* (6251) than by *Begonia* and *A. thaliana* (1565), which is expected given the closer taxonomic relationship *Begonia* has with *C. sativus* than *A. thaliana*. A larger number of genes are shared, or are conserved enough to be detected, between *B. conchifolia, A. thaliana* and *C. sativus* than by the latter two species and *B. plebeja*. This finding supports an elevated rate of genetic divergence in B. plebeja compared to *B. conchifolia*.

### 2.3.4 Functional annotation

To identify where key differences in expression lay between *B. conchifolia* and *B. plebeja* tissues, the annotations from the *B. conchifolia* genome were obtained for loci which were upregulated in one species or another by at least 5 fold. In petiole and leaf tissues, *B. conchifolia* appeared to have a higher number of sequences upregulated compared to *B. plebeja* (28 compared to 9, and 37 compared to 17, respectively). When comparing functions expressed in each tissue between species (appendices 4-15), *B. conchifolia* expressed genes related to defence response in each tissue. Interestingly, this trend was not found in *B. plebeja*; only one defence response gene was found to be upregulated in *B. plebeja* in vegetative bud. *B. conchifolia* leaf had an especially high number of defence response loci upregulated, 3 being annotated as involved in defence response, and 1 as defence response to bacterium. The higher number of genes expressed in this tissue in *B. conchifolia* compared to *B. plebeja*, and the composition of GO terms (enzymatic activity, protein kinase activity and metal ion binding) may reflect a concerted suite of loci involved in different aspects of response to stress. 7 unique loci were annotated as involved with defence response in *B. conchifolia* upregulated genes, indicating that the same genes are upregulated across tissues. *B. plebeja* showed many upregulated loci involved in photosynthesis, male flower and root being the only two tissues lacking loci annotated as being involved in photosynthesis and light harvesting. Interestingly, *B. conchifolia* shows an absence of upregulated loci directly involved with photosynthetic function in all tissues. While all precautions were taken to harvest material for RNA extraction at the same time and from tissues at the same stage, it cannot be ruled out that environmental effects at time of harvesting are responsible for the patterns seen. However, assuming that differences in environmental conditions and tissue

stages at harvest were minimal, these patterns suggest a fundamental difference in allocation of transcriptional activity between *B. conchifolia* and *B. plebeja*.

## 2.4 Discussion

### 2.4.1 Assembly strategy

Genomic resources for the mega diverse genus *Begonia* were further developed in this chapter by the sequencing of transcriptomes from six different tissues from two closely related but ecologically and morphologically distinct species of *Begonia, B. conchifolia* and *B. plebeja*. Further aims for the transcriptomic resources include using expression data to investigate how gene and genome duplication in *Begonia* may have contributed to its vast phenotypic diversity, and whether expressional divergence may show signatures of diversification in duplicate genes. This chapter aimed to outline the methods by which the transcriptomes were obtained and the methods used to generate expression data. 240 million paired end reads were produced on one lane of a HiSeq rapid v1 machine to produce 32 libraries consisting of 6 tissues across 2 species, with 3 replicates per species. A reference transcriptome assembly was produced for each species, aiming to maximise the completeness and accuracy of transcript reconstruction. Studies in yeast (Nagalakshmi et al. 2008) and mouse (Francis et al. 2013) have shown varying numbers of reads required to obtain a majority of all expressed genes (> 90%), the variation in large part originating from the difference in genome and splicing complexity. Francis et al. (2013) found from a study involving six different phyla that as little as 30 million reads are required to assemble a representative assembly, and a read count of over 60 million did not provide significant

returns of novel transcript discovery. Given the *Begonia* RNA-seq libraries collectively constitute 120 million reads per species, and cover the majority of the tissues found in each species, we can be confident of a satisfactory level of coverage of all the transcripts found in *B. conchifolia* and *B. plebeja*. A second consideration made when producing the reference transcriptome assemblies for each species was the assembly strategy. The *de novo* assembly software Trinity was chosen due to equal or better performance to other *de novo* assemblers (Vijay et al. 2012, Zhao et al. 2011). Despite Vijay et al. (2012) noting the markedly more contiguous and complete assemblies yielded by a mapping based strategy rather than a *de novo* one, a *de novo* strategy was used to reduce biases that could be introduced by differential taxonomic distances when mapping *B. conchifolia* and *B. plebeja* reads to the *B. conchifolia* genome. Testing different assembly strategies using Trinity revealed that the number of reads used in an assembly had marked effects on the quality of the assembly. Assembling reads from replicates individually and then merging the 16 assemblies from each species into one assembly per species using the tr2aacds pipeline yielded worse results than when a similar strategy was used with all reads from each tissue. The skew of the former strategy towards shorter sequences suggests an insufficient number of reads was present to successfully reconstruct full length transcripts. The substantially larger number of bases in the replicate assembly compared with the tissue assembly also indicates that the smaller number of reads in each assembly disrupts graph contiguity in Trinity, preventing it from collapsing all reads into the correct sequence, resulting in misassemblies. By far the least successful assembly strategy, however, was assembly of all reads for each species together. The size distribution of assembled transcripts indicated a very high degree of fragmentation, and the number of bases, similar to the replicates assembly, was very high, suggesting a high rate of misassemblies and failure to collapse reads into the correct predicted transcript. Intuitively, a greater number of reads would appear to provide more information, and

therefore a better assembly. Previous work has suggested that, as transcriptome complexity increases, quality metrics such as the percentage of full length reconstructed transcripts decreases dramatically (Chang, Wang and Li, 2014). The probability of misassembly increases with greater number of reads, possibly due to this increase in complexity (Conesa et al. 2016). All reads together may have provided an overwhelmingly large number of isoforms which were not easily traversed by Trinity's algorithm, resulting in an assembly which failed to accurately explain the reads. The tissue transcriptome assembly strategy, therefore, may act to compartmentalise this complexity within each individual tissue assembly, while still containing enough reads to accurately represent the isoform complexity within each tissue. The ability to assemble transcriptomes from reads originating from each tissue sequenced and then merge the resulting assemblies into one assembly representing all tissues was facilitated by the EvidentialGene pipeline tr2aacds. Particularly important was the correct identification and removal of redundant sequences in the merged assembly. The demonstration of the utility of EvidentialGene in this chapter is supplemented by other studies (Visser et al. 2015, Chen et al. 2015, Postnikova et al. 2015); the observation that differences in algorithm and parameters between available assemblers suggests that no one software will be able to fully reconstruct every transcript. Thus, the ability to use a number of assemblers with each one's own parameter space to reconstruct a transcriptome, and subsequently merge all the resulting assemblies while avoiding issues of redundancy is very useful when attempting to reconstruct the optimum transcriptome (Nakasugi et al. 2014).

2.4.2 Differences in expression patterns between *B. conchifolia* and *B. plebeja*

The number of expressed genes per tissue in *Begonia* was generally similar between the two species, both having an elevated number of genes expressed in female flowers, petioles, vegetative buds and roots (Table 2). Between 5 and 10% of the total genes in *Begonia conchifolia* were found to be expressed in any one tissue, the highest being *B. conchifolia* female flower, expressing 8.7% of the total 21856 transcripts in the *B. conchifolia* genome predicted transcripts, while *B. plebeja* leaf showed the least expression, with 6.1%. Though upwards of 95% of reads from each library were mapped to the *B. conchifolia* predicted transcripts by Salmon, many loci were eliminated by EdgeR due to insufficient expression in enough treatments. This step, while eliminating much data, allows remaining data to be analysed in a statistically robust manner. The tissue expressing the smallest number of transcripts was leaf, followed closely by male flower. A less diverse repertoire of genes in male flowers is supported by similar results found in *P. sativum* (Alves-Carvalho et al. 2015), where expressed gene numbers in flower are in the minority. This is surprising, given the developmental complexity of male flowers and the diverse number of substructures they contain. The sequencing of whole flower tissue may have masked the expressional diversity of the more complex substructures within *Begonia* male flowers; a study in *Vitis vinifera* (Fasoli et al. 2012) showed that splitting out of different flower components in the analysis of transcriptional diversity revealed marked differences in number of genes expressed between stamen, pollen, petal and carpel. The differences in number of genes expressed in flowers between different species may originate from the heterogeneity of both sexual systems and the underlying genetic pathways that encode them. While some studies have suggested that a common suite of genes is involved in the manifestation of different mating systems (Guo et al. 2010), other studies have suggested that a great deal of variation is present in the genetic

basis of different mating system shifts (Fishman et al. 2002, Slotte et al. 2012, Zuellig et al. 2014). The difference in number of expressed genes in male flower and female flower may reflect divergent complexity between the two flower types. Interestingly, a number of studies have shown that newly duplicated genes tend to be preferentially expressed at first in pollen before they are co-opted into functions in other tissues (Wang et al. 2016, Cui et al 2015). This observation may underlie the lower number of transcripts expressed at a significant level in male flower; while more transcripts may be expressed, some of these may be newly duplicated genes, which have been found to be downregulated in early stages of evolution (Adams et al. 2005), therefore these may be missed when identifying genes significantly expressed in *Begonia*. Male flower has more differentially expressed genes between *B. conchifolia* and *B. plebeja* compared to the other tissues sampled; 695 genes (3.2% of total *B. conchifolia* genes) were downregulated, while 753 (3.4%) were upregulated.

While female flower, leaf, petiole and vegetative bud all had similar numbers of genes upregulated and downregulated in both *B. conchifolia* and in *B. plebeja*, root tissue had the most conserved expression, with the fewest number of genes differentially expressed between the two species. This finding is not surprising in light of the selective pressures acting on *Begonia* beneath ground; a large proportion of *Begonia* species are found growing on limestone rockfaces or in limestone rich soils, this soil preference seeming to be widespread across African, South East Asian and South American *Begonia*s. Roots also had the highest proportion of tissue specific transcripts, with 1247*(B. conchifolia)* and 1334 (*B. plebeja)* transcripts being expressed exclusively in roots. Previous studies have noted high proportions of transcripts being expressed specifically in root or root derived tissue such as root nodules (Verdier et al. 2013).

Annotation of genes upregulated in *B. conchifolia* compared to *B. plebeja* and genes upregulated in *B. plebeja* compared to *B. conchifolia* revealed photosynthetic genes to be upregulated and stress response genes to be downregulated in *B. plebeja*. A pattern of increased photosynthetic activity and suppressed response to stress has been found in *A. thaliana* seedlings in response to increasing competition from neighbouring plants (Geisler et al. 2012), suggesting that as resources such as water and light become more limited, plants respond by allocating more resources to growth and less to stress response. Studies in rice have shown upregulation of photosynthetic genes in response to drought. Increased accumulation of sugars via increased photosynthetic rate may allow leaves to develop a more negative osmotic potential, preventing excess stress in drought conditions by maintaining turgidity in leaves (Thalmann et al. 2016, Ambavaram et al. 2014, Todaka et al. 2015). Though this has been found in a number of studies involving rice, other studies have found the opposite effect of drought stress on regulation of photosynthetic genes (Degenkolbe et al. 2009), therefore the relationship between drought response and photosynthesis is yet to be fully clarified. Drought stress features differently in the habitats of *B. conchifolia* and *B. plebeja*, the former occupies moist shaded understorey, whereas the latter is distributed in more open, dry tropical forests. The finding that *B. plebeja* displays signatures of increased photosynthesis and reduced stress response may suggest an adaptive response to more widespread conditions of drought, putting more resources into growth and providing more solutes to optimise tissue osmotic potential. This hypothesis will require further validation, however, as it is possible that transient changes in stress levels between the two plants at harvesting has led to artificially different responses.

### 2.4.3 Conclusions

The generation of RNA-seq data for *B. conchifolia* and *B. plebeja* has expanded the resources available for the genus, providing more opportunities for qualitative and quantitative studies in this group. The availability of the *B. conchifolia* genome provides more information about both coding and non-coding sequence in *Begonia*, however the genus is still very much a non-model species. The addition of transcriptomic data for species such as *B. plebeja* for which no genome is available allows the rapid generation of large amounts of coding sequence data without the computational overheads associated with assembling whole genome data. Development of novel methods for assembly and analysis of transcriptome data has facilitated the increase in sequence data available for previously understudied organisms where a reference genome is unavailable (Kannan et al. 2016, Patro et al. 2015). Aside from the immediate benefits of the faster and cheaper access to a large amount of sequence data, the quantitative quantitative data for genes in the form of expression data RNA-seq provides gives context to data not available from nucleotide sequence alone. Such data is especially important in the dissection of species differences; much phenotypic change has been ascribed to changes in expression pattern not necessarily accompanied by a significant change in coding sequence (Meunier et al. 2013, Sakuma et al. 2013). Therefore, whole genome data in *Begonia* is important when attempting to understand the basis of the high phenotypic and ecological diversity seen across the genus, it must be supplemented by quantitative expression data to fully understand the basis of diversity.

# Chapter 3: Identification of candidate genes putatively underlying diversification in *Begonia*

## 3.1 Introduction

### 3.1.1 Duplication and diversity

Gene and genome duplication is a major force in the evolution of biological complexity. As a pervasive phenomenon, duplication has been a feature of evolution of most forms of life to various extents. Mammalian genomes undergo segmental and whole genome duplication relatively infrequently due to high dosage sensitivity and tightly regulated development (Leitch and Leitch, 2008). A number of whole genome duplications have been detected in fish. Though there is contention regarding the date of the duplications, it is widely believed that duplicated genes facilitated the rapid diversification of the ray-finned fishes during the Jurassic and Cretaceous periods (Christoffels et al. 2004).Of all the kingdoms, however, the group which has the most extensive history of duplication is the plants. Using age distributions of duplicate genes, both ancient and recent WGDs have been found across the angiosperms. Taking into account ancient WGDs in basal relatives of angiosperms, some species have in their evolutionary history as many as four WGDs, including species such as potato, wheat and cotton (Adams and Wendel, 2005). Some of the astonishing phenotypic and high rates of speciation found in the angiosperms has been attributed to the effects of duplication. Following the doubling of genetic material in a WGD event, genomic events such as genomic shock are commonly activated, involving processes such as elevated transposable element activity and genome downsizing (McClintock, 1983). Genome downsizing involves the deletion of large portions of the genome, as well as chromosomal

breakages and fusions; many palaeopolyploids behave functionally as diploids as a result of genome downsizing (Bennett, 2004). For example, Bowers et al. (2003) found that only 30% of *A. thaliana* genes originating from a duplication ~85 MYA were retained. The authors also noted the contrast to rates of mammalian genome evolution; comparing mouse and human genomes showed the two genomes to be 70% syntenic (Chinwalla et al. 2002), suggesting that the rate of genome evolution, and downsizing, in plants may be very rapid. Evidence from *Tragopogon* also suggests that, in addition to post duplication changes being rapid, they are predictable and repeatable, suggesting that deletion of different parts of the genome is guided in part by selective forces (Buggs et al. 2012) (though see Weiss-Schneeweiss et al. 2011). The outcome of these post duplication processes is that new derived phenotypes are able to be produced and exposed to selection, enabling the evolution of new traits, and potentially new species (Puttick et al. 2015, Adams and Wendel, 2005). Following a duplication event, the fate of the duplicated genes is dependent upon the mechanism of duplication; Freeling (2009) posits that the extent of interaction between genes biases their probability of retention – if a gene interacts frequently with other genes, for example in multi-gene complexes. Patterns of gene loss are found across plant taxa, certain genes, often housekeeping genes, are consistently found in single copy number regardless of different duplications histories (Paterson et al. 2006). Similarly consistent patterns are seen in multigene families; some functional groups have a greater propensity to be retained after segmental or whole genome duplication. Groups frequently found to be retained include genes involved with regulation and response to bacteria and environmental stimuli.
Such results support a functional context for duplicate gene fates; roles of genes, their mode of action and regulation, and the presence of pleiotropy and epistatic effects all contribute to the combined selective pressure for either gene loss or retention.

### 3.1.2 Duplication in the mega-diverse genus *Begonia*

*Begonia* also has an extensive history of genome duplication and rearrangement, and shows signs of a rapidly evolving genome, with chromosome numbers ranging between 16 and 156 (Neale et al. 2006). Surveys of C-value and chromosome number across the sections of *Begonia* have shown that high variation is seen both between and within sections, and changes in chromosome number indicate that both whole genome duplications and smaller duplications of chromosomes and large genome segments are frequent (Dewitte et al. 2009) (Figure 1). This finding is supported by cytological studies, showing that meiosis is often aberrant in hybrids, the formation of univalents and multivalents causing unbalanced chromosome numbers, though interestingly hybrids are often fertile (Dewitte et al. 2010).

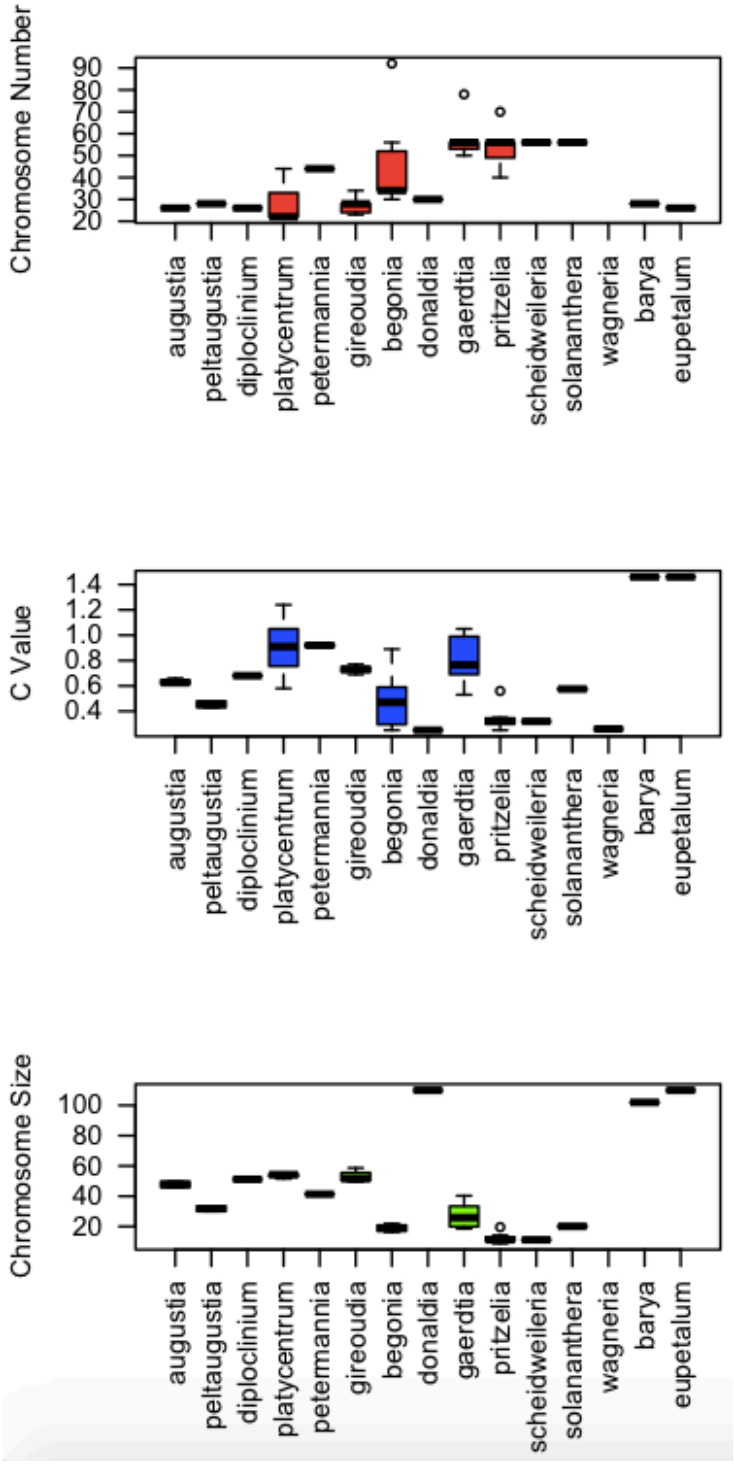



Figure 1. Figure made from data in Dewitte et al. 2009. Chromosome number (red), C-value (blue), and chromosome size (green) are plotted in 15 sections of *Begonia*

Though polyploidy and gene duplication has gained much attention in the study of *Begonia*, the focus has mainly been on large scale patterns of chromosome size, as well as studies of diploid gamete production in a range of *Begonia* species. Based on the prevalence of duplication in the genus, a study of how duplication has impacted *Begonia* evolution is much needed. Answering key questions such as what has been the fate of duplicate genes in different species in terms of expression pattern and sequence divergence will be vital to understanding to what extent *Begonia* owes its phenotypic diversity and ecological success to it's highly dynamic genome.

### 3.1.3 Genomic resources available for studying diversity in *Begonia*

The availability of genomic resources in *Begonia* enables evolutionary studies in the genus. A draft genome assembly is available for *B. conchifolia* (Chapter 2). Annotation of the *B. conchifolia* genome provides a set of predicted transcripts derived from the genome, permitting the use of all the coding sequence in the genome for sequence based analyses. To obtain quantitative data for duplicate genes, RNA-seq was used to generate expression data for *B. conchifolia* and the closely related but ecologically and phenotypically divergent species, *B. plebeja* (Chapter 2). Libraries from six tissues (female flower, leaf, male flower, petiole, root and vegetative bud) with three replicates per tissue were sequenced at Edinburgh Genomics on a HiSeq rapid v1 machine, generating ~240 million 150bp paired-end reads. Assemblies for *B. conchifolia* and *B. plebeja* yielded around 20,000 unique loci, with an N50 of around 2000 for both assemblies and sequence lengths ranging from 300bp to ~15,500bp.

### 3.1.4 Premise of the chapter

A key aim of this chapter is to understand the duplication landscape of *Begonia*, and specifically ask what is the fate of duplicate genes in terms of their expression profile. Understanding how much, to what extent, and in which functional categories duplication has resulted in divergent expression patterning between *B. conchifolia* and *B. plebeja* will shed light on why these two closely related species are so phenotypically and ecologically divergent. In a wider context, this will also help to understand which ecological pressures have been important in driving *Begonia* diversification, and ultimately, what factors have made *Begonia* such an ecologically successful genus. The predicted transcripts from the *B. conchifolia* draft genome and the primary transcripts of three outgroup species, *C. sativus*, *A. thaliana* and *G. max* will be grouped into gene families to investigate duplication patterns in *Begonia* in a wider taxonomic context. Duplicate genes in *B. conchifolia* across all gene families will be taken in a pairwise fashion, and expression patterns between each pair will be analysed. To do this, RNA-seq reads from six tissues from *B. conchifolia* and *B. plebeja* will be mapped to each duplicate pair, and the expression divergence of *B. conchifolia* duplicates will be compared to the expression divergence of *B. plebeja* duplicates. If gene or genome duplication has facilitated the phenotypic and ecological divergence of *B. conchifolia* and *B. plebeja* via the divergence of expression patterns between duplicate genes, it would be expected that there are duplicate genes which have diverged in expression pattern in one species but not the other. Rather than divergence of expression between duplicates in both species, which may simply suggest a neutral partitioning in expression over time, divergence in one species and not the other indicates the change in expression may be adaptive and specific to a species. The contrasting habitats *B. conchifolia* and *B. plebeja* are found in and their divergent phenotypes suggest that different environmental challenges may be a driver of

their divergence. If this is the case, we would expect genes which are found to be divergent in one species but not another to be enriched in functions which are involved with environmental responses.

## 3.2 Methods

### 3.2.1 Clustering sequences into gene families

*Begonia conchifolia* transcripts were grouped into gene families along with *Cucumis sativus, Glycine max and Arabidopsis thaliana* primary transcripts. Predicted primary transcripts were used for *C. sativus (Cucumis sativus* v1.0, no reference publication available), *G. max* (*Glycine max* Wm82.a2.v1, Schmutz et al. 2010) and *A. thaliana* (*Arabidopsis thaliana* TAIR10, Lamesch et al. 2012). After examining the *B. conchifolia* transcripts predicted from the genome, this sequence dataset was deemed to be the most complete out of all the genomic resources available, but some sequence duplication was identified, where exact copies of the same sequence appeared twice. To prevent this from artificially inflating estimates of copy number, the duplicate sequences were eliminated using the tr2aacds pipeline within the EvidentialGene package (Gilbert, 2013). The pipeline uses CDS and peptide translations of the input transcripts to assign them to a filter pass set, a set of putative isoforms, and a dropped set, which are excluded from further analyses.

The final sets of *Begonia* and outgroup transcripts were clustered using OrthoFinder (Emms and Kelly, 2015), chosen due to better performance compared to many other clustering methods, and because of the elimination of bias introduced by both sequence length and differential taxonomic distances between input species. Homologous sequences were identified from the clusters formed by OrthoFinder. The inability to guarantee the completeness of the transcriptomes means that incomplete sampling may lead in some cases to artificial patterns of gene gain or loss, and may result in incorrect out-paralog assignment. While it is impossible to correct this, the bias that may be introduced is kept in mind when interpreting results.

### 3.2.2 Estimating *Begonia* expression divergence

Expression divergence was estimated both between and within *B. conchifolia* and *B. plebeja*. Reads from tissue replicates were mapped individually to *B. conchifolia* predicted transcripts using Salmon 0.7.2 (Patro, 2015) with default parameters, and raw counts from all used tissue replicates were used to construct a matrix. EdgeR 3.16.5 (Robinson et al. 2010) was used to calculate relative expression using the TMM method in order to ensure that expression was normalised between samples. The relative expression values were used to calculate an average of relative expression for each tissue across all replicates present. Average relative expression of each predicted transcript was therefore available for all tissues in *B. conchifolia* and *B. plebeja*. To compare expression divergence for a predicted transcript, the Pearson correlation coefficient was calculated for relative expression within *B. conchifolia*, within *B. plebeja*, and between *B. conchifolia* and *B. plebeja*. The EdgeR analysis retained transcripts which had at a counts per million value for at least 100 in at least two samples, thus eliminating many transcripts which had low expression. Therefore, the correlation coefficient of relative expression was calculated for transcripts with available data, however many transcripts lack sufficient expression to calculate this metric robustly.

### 3.2.3 Identifying *A. thaliana* orthologs

Though *A. thaliana* is one of the species included in the gene family clustering, it is impossible to guarantee the presence of an *A. thaliana* ortholog in each gene family due to transcriptome incompleteness, lineage specific gene families or gene loss. The presence of an *A. thaliana* ortholog is required for confident translational alignment; due the draft status of the genome and lack of annotation curation, it is not certain whether *Begonia* predicted

transcripts are full length or contain both start and stop codons. Therefore, a robust way of inferring the correct reading frame and peptide sequence in this case is by alignment to an *A. thaliana* ortholog. The peptide sequence of the ortholog or orthologs in *A. thaliana* are identified with BLASTX 2.2.3 (Altschul et al. 1990), using all sequences in the gene family as a query. An E value threshold of $1e^{-40}$ is used, and the maximum target sequences reported is set to 1. The results are then parsed to eliminate hits which have less than 40% identity and have a length of at least 200 base pairs. Gene families continue to downstream analyses only if an ortholog is identified in *A. thaliana*. Such stringent thresholds are in place due to the importance of the ortholog to the translational alignment step.

### 3.2.4 Translational alignment

Sequences within each gene family were translationally aligned to maximise the quality of the alignment and the robustness of downstream analyses based on the alignment.
For each gene family, the longest open reading frame (ORF) was found for each sequence within, defining an ORF as the longest length of sequence without a stop codon in any of the 6 reading frames. The *A. thaliana* ortholog is aligned with the gene family members using MAFFT v6 (Katoh et al. 2002), using the '--adjustdirectionaccurately' flag, gaps and the *A. thaliana* ortholog are removed from the subsequent alignment. This step allows all sequences to be in the same orientation, which is convenient for executing downstream steps. *A. thaliana* alone is used for outgroup homology; the purpose of the sequence is for reading frame assessment purposes only and wider taxon sampling will needlessly increase alignment complexity. At this step, all sequences shorter than 300 base pairs are removed, preventing short sequences from introducing error in alignments and further analyses. To produce a protein alignment of the gene family members, the nucleotide sequences are translated into

all six reading frame using dna2pep 1.1 (Wernersson, 2006), and aligned with the *A. thaliana* protein ortholog using MAFFT. A distance matrix is made with the alignment using clustal omega v2 (Sievers and Higgins, 2014), and for each member of the gene family, the reading frame translation with the lowest distance to the *A. thaliana* protein ortholog is taken as the correct reading frame for the gene family member. Once the correct reading frame is identified, the peptide sequences are aligned using MAFFT. The nucleotide sequence is edited based on the correct orientation; either none, one or two bases are removed from the beginning or end of the sequence depending on the reading frame.
Pal2Nal v14 (Suyama et al. 2006) is used to perform the translational alignment using the gene family peptide alignment and nucleotide sequences using default parameters.

### 3.2.5 Estimating *Begonia* pairwise sequence divergence

Pairwise divergence is calculated between all sequences in the gene family, both within *Begonia* and between species. Distmat from the EMBOSS package suite (v2.7.1) (Rice et al. 2000) was used to calculate the distance matrix with the Jukes-Cantor distance correction method.

### 3.2.6 Functional annotation

Gene families were annotated by querying *A. thaliana* peptides with all gene family sequences using BLASTX 2.2.3 at an evalue threshold of $1e^{-40}$. Only gene families with a hit at this threshold were used in further analyses.

### 3.2.7 GO enrichment analysis

Candidate gene families identified were used to test for functional enrichment. The online GO term enrichment analysis tool from AgriGO (Du et al. 2010) was used to perform all analyses. The closest relative to *Begonia* with a reference genome was used as background reference. All *Begonia* candidate genes were used in a BLASTX search against cucumber pepetides, and hits with at least 70% identity across at least 200 base pairs. The identified *C. sativus* genes were then used as input, using the *C. sativus* background reference available on AgriGO.

### 3.2.8 Phylogenetic Analysis

A gene tree was drawn of each identified candidate gene families, with the inclusion of additional taxa and outgroups. Additional taxa used were *A. thaliana*, *C. sativus* and *G. max*, ensuring adequate taxon sampling. *Amborella trichopoda* was used as an outgroup for all candidate gene trees. Identification of additional taxa and outgroup species homologs was done using BLASTX, using no E value threshold. Top BLAST hits were aligned together with *Begonia* sequences and the alignment was manually inspected for homology across the coding sequence (eliminating artefacts resulting from domain homology).
Phylogenetic analysis of candidate genes was undertaken using MrBayes 3.2 (Ronquist and Huelsenbeck, 2003), using the GTR model. The analysis was run using two chains under default settings for 10,000 generations, sampling every 10 steps. Additional repeats of 10,000 generations were added until the standard deviation of split frequencies was below 0.01.

## 3.3 Results

### 3.3.1 Duplication patterns in *Begonia*

Based on previous observations that *Begonia* has undergone a genome duplication, patterns of duplication, and how these relate to both sequence divergence and expression diversity were investigated. The distribution of *B. conchifolia* gene family copy numbers was compared to the other outgroup species, *C. sativus*, *A. thaliana* and *G. max* to investigate how duplicated the *Begonia* genome is in a wider taxonomic context.

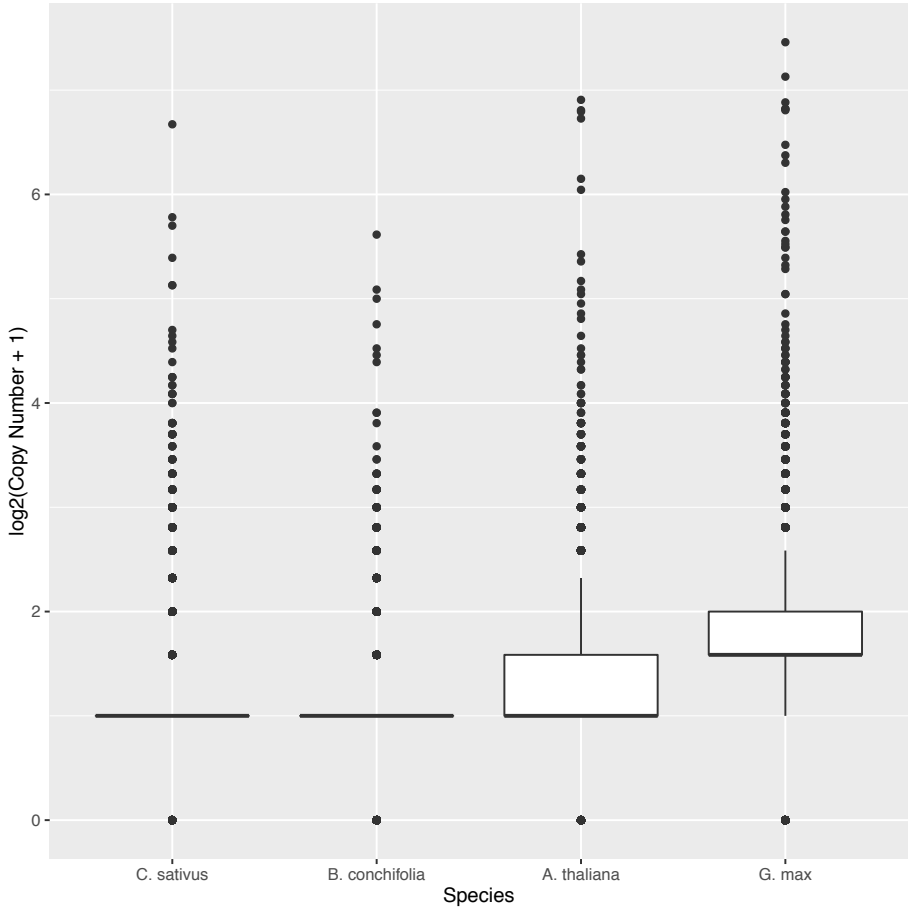



Figure 2. Number of genes across gene families in *B. conchifolia, C. sativus, A. thaliana* and *G. max*. 1 is added to each copy number to eliminate copy numbers of 0, allowing $\log^{2-}$ transformation.

Figure 2 shows copy number values which have have 1 added to each entry point, allowing $\log^2$ transformation for better visualization. Statistics discussed further in the text refer to untransformed values.

Comparison of gene family sizes between *B. conchifolia*, *A. thaliana*, *C. sativus* and *G. max* showed that *G. max* had the highest frequency of duplications among gene families (Figure 2). *G. max* has a median copy number of 2, compared to a median of 1 for each of the other species, as well as a mean copy number of 2.662, reflecting the large number of genes that are duplicated and still retained. The frequency of duplicated gene families corroborates extensive evidence of genome duplication in *G. max* (Clarindo et al. 2007).

*A. thaliana*, like *G. max*, has had an extensive history of duplication, with reports of between one and three duplication events happening at various intervals in *A. thaliana*'s evolutionary history (Simillion et al. 2002). Contrary to *G. max* and *A. thaliana*, *C. sativus* has been reported to have relatively few duplicated gene families, owing to the absence of a recent whole genome duplication (Huang et al. 2009). This is reflected in the gene family sizes of *C. sativus*, with a mean gene family size of 1.328 compared to 1.512 for *A. thaliana*.

Compared to the three outgroup species, the *B. conchifolia* genome has a lower duplication rate, with a mean gene family size of 1.114, and a maximum gene family size of 48; less than half of each of the outgroups. The larger number of genome duplications in *A. thaliana* and *G. max* may underlie the difference in duplication patterns between *B. conchifolia* and the former two species; genome downsizing may have eradicated most signals of a single, ancient, large scale genome duplication in *Begonia*. The issues of incompleteness introduced by the draft *B. conchifolia* genome, however, cannot be discounted, and the contribution of this towards an apparently lower duplicate gene content in *Begonia* must be first established.

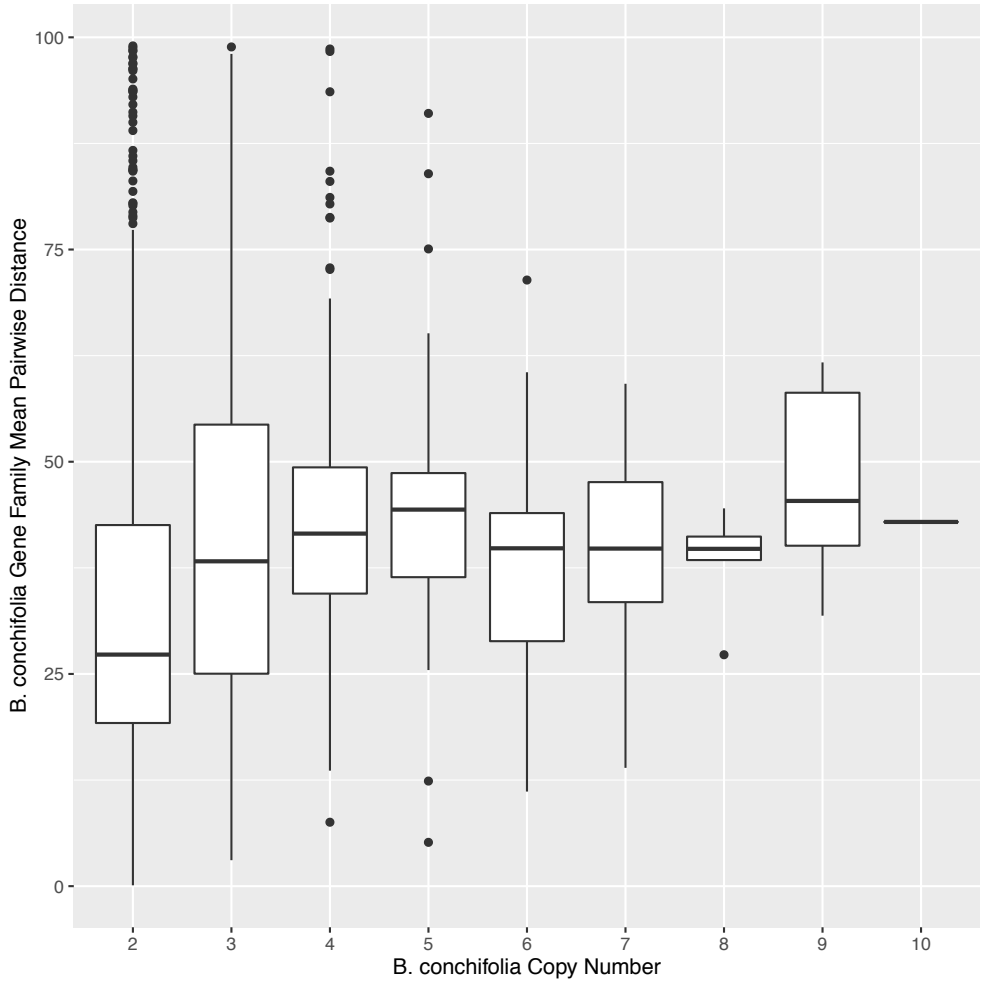



Figure 3. Mean pairwise sequence distance of gene families across a range of gene family sizes.

The mean pairwise divergence between duplicate pairs increases with gene family copy number in *B. conchifolia*, though this trend is only evident in smaller gene families (copy numbers between 2 and 5) (Figure 3). Distinct groups are not formed by gene families of different sizes; much overlap is seen between groups. Gene families larger than five members do not show a continuing increase in divergence between family members, median divergence remaining roughly equal at around 38%. No drop is seen in mean pairwise divergence in larger gene families, the lower bounds of mean pairwise divergence are consistently higher in gene families with more than 5 members, the majority of these families having a mean pairwise divergence not lower than 12.5%. This trend, which is mirrored by the outgroup species, is indicative of a mutational threshold beyond which it is likely that mutations will start to cause loss of function mutations or lead to premature stop codons in the protein in question.

### 3.3.2 Expression patterns in *Begonia*

No relationship was found between expression divergence between a gene pair and the mean pairwise divergence of the gene family the gene pair is placed in. Figure 4 shows the mean pairwise divergence plotted against expression divergence per gene family in *B. conchifolia*. The same graph for *B. plebeja* shows a similar trend and is not shown here. The bulk of gene pairs have a sequence divergence of between 10% and 50%, a small proportion of outliers having a much greater mean sequence divergence. This variation, however, does not in any way appear to have an association with expression divergence in either species (*B. conchifolia* Pearson Correlation Coefficient = -0.052, *B. plebeja* Pearson Correlation Coefficient = -0.048).

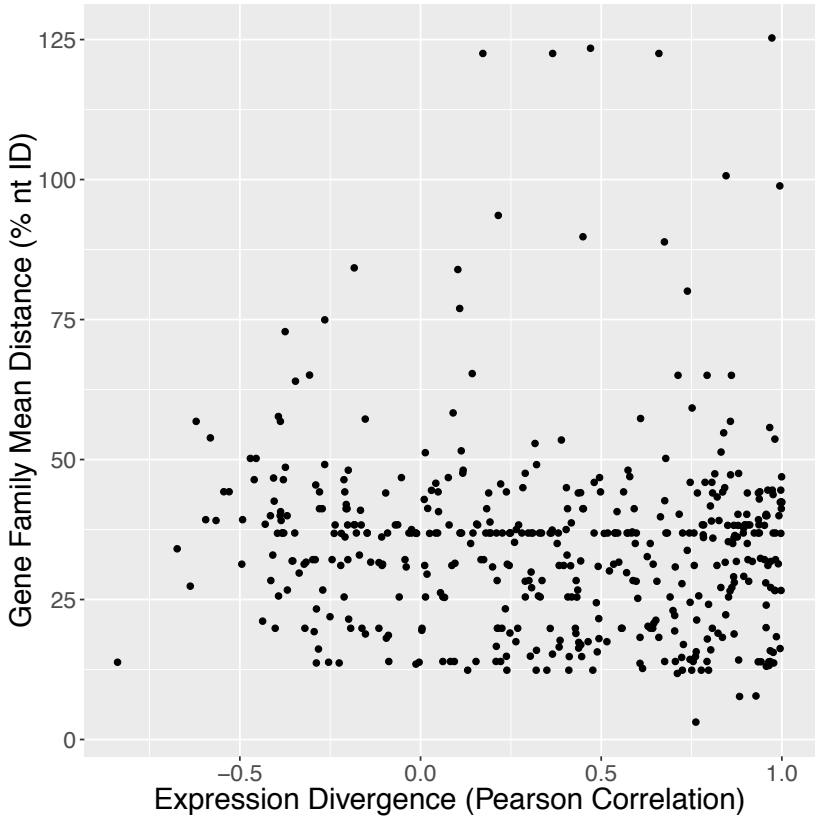



Figure 4. Expression divergence of *B. conchifolia* duplicate gene pairs plotted against the mean pairwise sequence distance of the gene family from which the duplicate gene pair originated.

When partitioning gene families by families with fewer than 3 members and greater than 2 members, a significant difference was seen between expression correlation of gene pairs within gene families in these two size bins (Figure 5). The difference was significant at an alpha of 0.05 between *B. conchifolia* gene families of size less than 3 and families more than 2 (Wilcoxon rank sum test W = 21883, P-value = 0.01), and was also significant between *B. conchifolia* gene families of size less than 3 and *B. plebeja* families of size more than 2 (Wilcoxon rank sum test W = 21516, P = 0.02). Comparing families of size less than 3

between *B. conchifolia* and *B. plebeja* did not show a significant difference (Wilcoxon rank sum test W = 5205, P = 0.99), and similar results were found when comparing gene families of size greater than 2 between *B. conchifolia* and *B. plebeja* (Wilcoxon rank sum test W = 66843, P = 0.76).

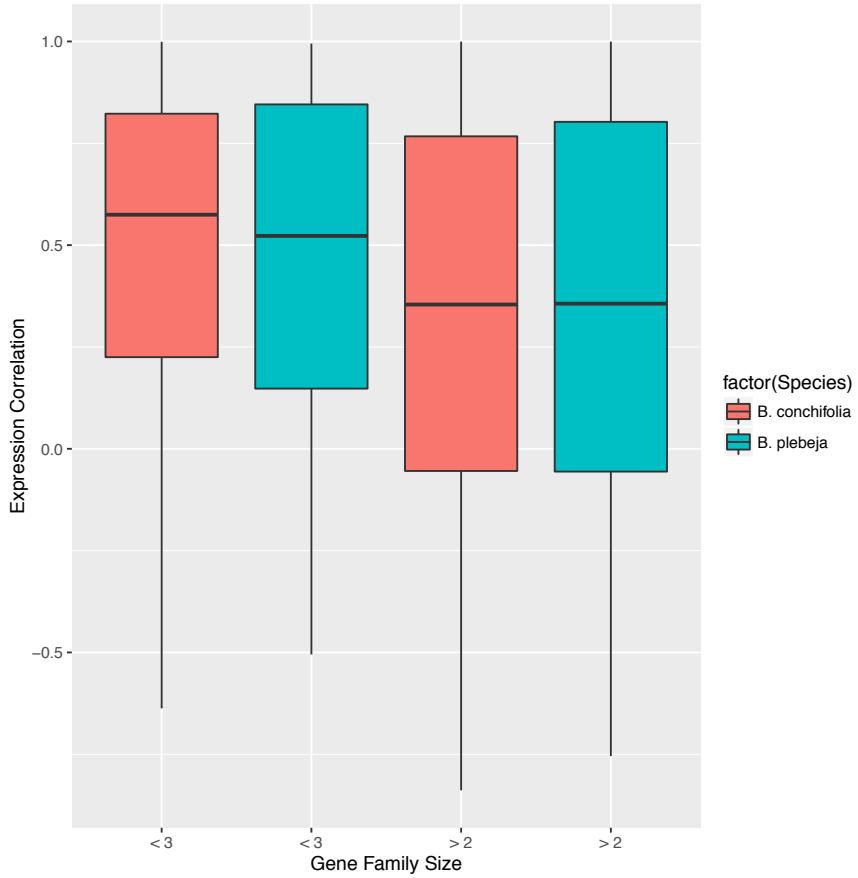



Figure 5. Expression correlation of duplicate gene pairs in *B. conchifolia* and *B.plebeja* in gene family sizes of less than 3 and greater than 2.

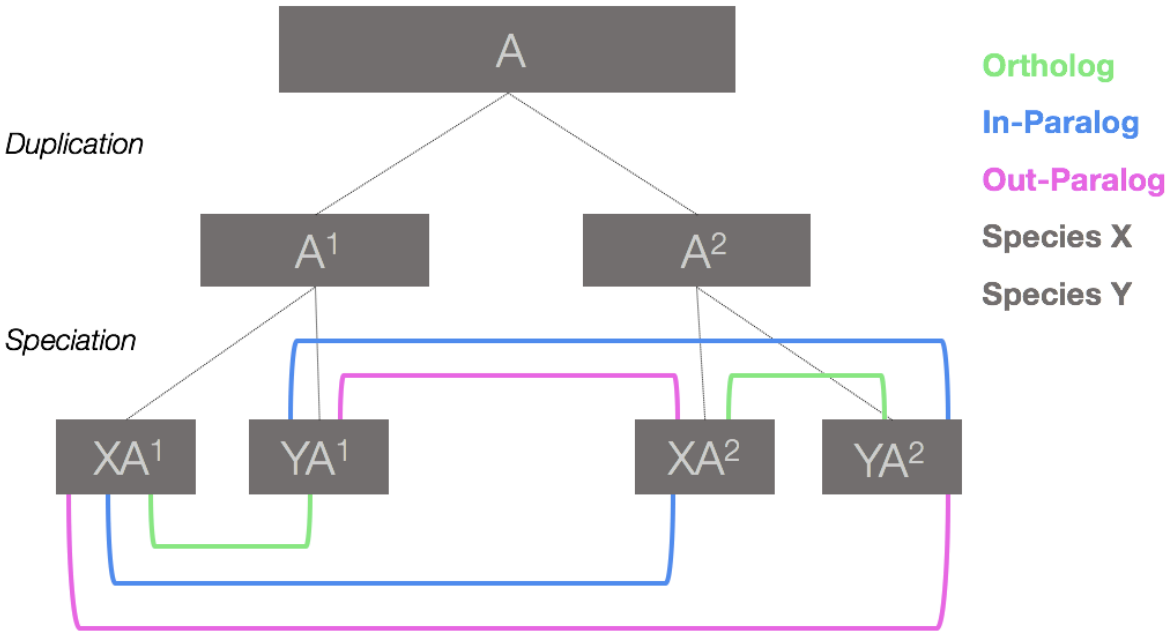



Figure 6. Duplicate gene relationship nomenclature between and within species.

Using gene families where at least one duplication had occurred in *B. conchifolia*, expression values were compared across a number of different homologous relationships. The terminology used here and illustrated in figure 6 and figure 7 is based on definitions from CoGe, a plant comparative genomics framework (Lyons and Freeling, 2008).

Figure 6 shows the relationship of a duplicate gene pair in two different species. An ancestral gene 'A' undergoes a duplication event in an ancestral species, producing two paralogs, $A^1$ and $A^2$. Following this, a speciation event occurs, where the ancestral species, containing two copies of ancestral gene A, undergoes lineage divergence to form two species, X and Y. The duplicate genes are now referred to as $XA^1$ and $XA^2$ in species X, and $YA^1$ and $YA^2$ is species Y. The duplicate $A^1$ in each of the two species $XA^1$ and $YA^1$ are referred to as orthologs, as are $A^2$ in $XA^2$ and $YA^2$. Duplicate genes which are located in a single species, $XA^1$ and $XA^2$, as well as $YA^1$ and $YA^2$ are referred to as in-paralogs. Genes derived from the

duplication event and are also located in different species, in this example the pairs $XA^1$ and $YA^2$, and $XA^2$ and $YA^1$, are referred to as out-paralogs.

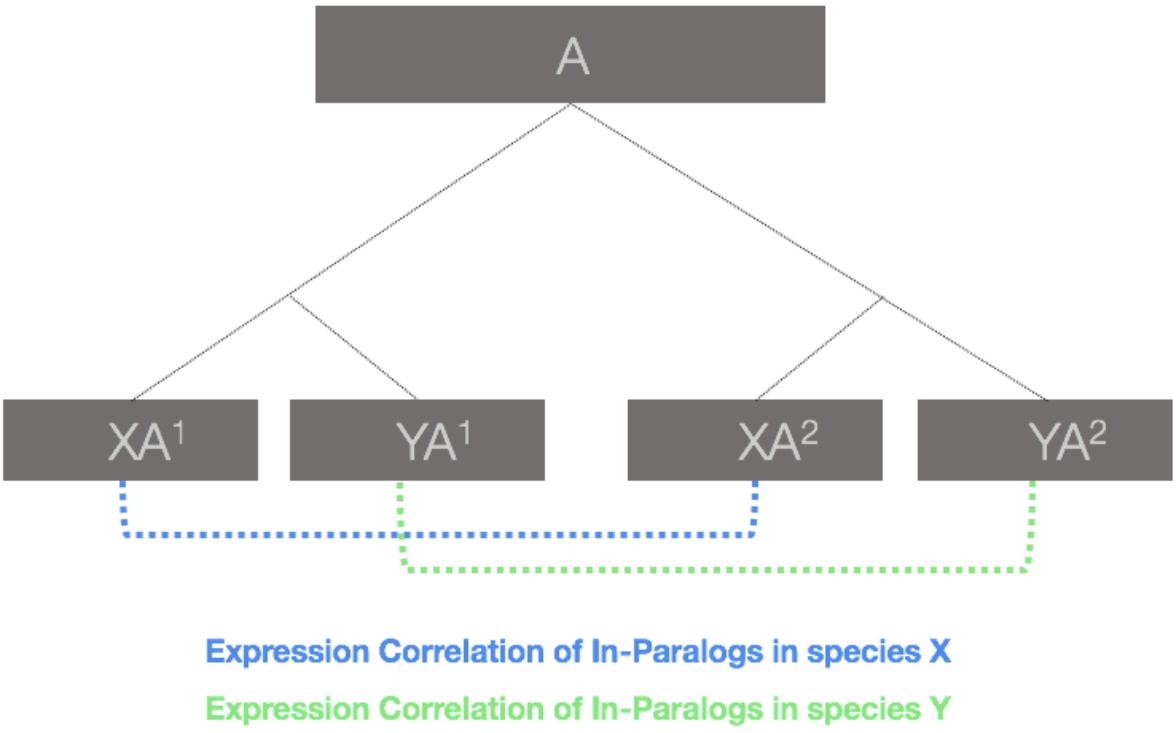



Figure 7. Expression correlation calculated between in-paralogs between two species

To investigate whether gene duplication may have caused divergence in expression patterns between *B. conchifolia* and *B. plebeja*, the expression correlation between in-paralogs in *B. conchifolia* was compared to expression correlation of in-paralogs in *B. plebeja* (Figure 8). Relative expression values were calculated using EdgeR (Robinson et al. 2010), using TMM as the method of normalisation. Expression correlation values were obtained from across identified gene families, where each *Begonia* gene family member was compared to every other *Begonia* gene family member in a pairwise fashion. For each gene family member, the

mean normalized relative expression value of the replicates was taken for each tissue, resulting in one estimate of relative expression for each tissue in each species.

A Pearson correlation coefficient was calculated for each pair of genes across the six tissues sampled, using in-paralogous expression values only, thus comparing *B. conchifolia* expression values in one duplicate to *B. conchifolia* expression values in the second duplicate. This is repeated for *B. plebeja*, resulting in, for one gene duplicate pair, the correlation of expression between the duplicates in *B. conchifolia* and in *B. plebeja*. In more expanded gene families, this is repeated for each pairwise comparison on *Begonia*.

The methods implemented in EdgeR for normalisation require that each locus must have high enough expression across tissues in order to have sufficient statistical power to calculate relative expression. Due to low expression in a large number of genes, expression correlation was obtained from as many gene pairs as possible, however, many pairwise comparisons have no expression correlation data. All available pairwise comparisons of expression correlation across gene families in *B. conchifolia* in-paralogs were plotted against the corresponding expression correlation in *B. plebeja* in-paralogs (Figure 8).

Each point represents one gene duplicate pair, with the position along the Y axis corresponding to the similarity of expression in *B. plebeja* in-paralogs, and the similarity of expression in *B. conchifolia* in-paralogs along the X axis in the same pair. A value of 1 indicates a perfect positive correlation of expression across tissues, and values further towards 0 indicate a lower similarity of expression across tissues.

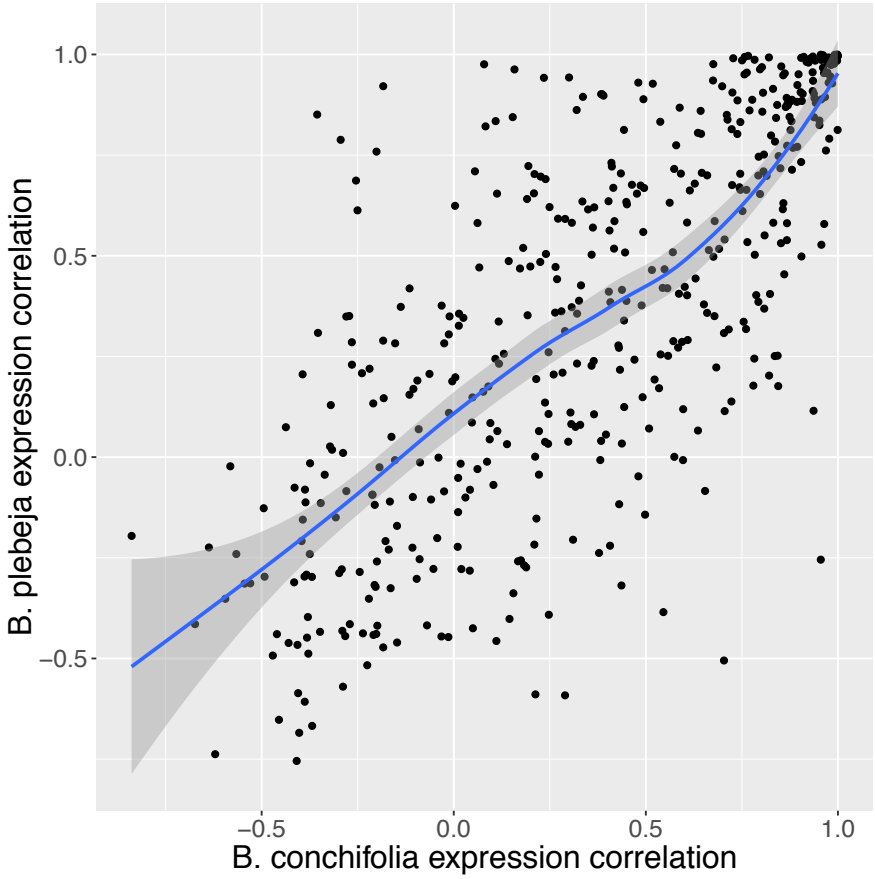



Figure 8. Expression correlation between in-paralogs of *B. conchifolia* plotted against the expression correlation between homologous in-paralogs of *B. plebeja*. Fitted line uses Loess smoothing, width indicates 95% confidence intervals.

The association of *B. conchifolia* expression correlation and *B. plebeja* expression correlation shows a linear distribution, with a positive correlation (Pearson Correlation Coefficient of 0.75, N=470). A cluster of 31 gene pairs where expression correlation in both species is more than 0.95 indicates a considerable number of duplicate pairs have conserved expression within species. Around a third (32%) of gene pairs had expression profiles which were equally conserved between *B. conchifolia* in-paralogs and *B. plebeja* in-paralogs, this group being defined by *B. conchifolia* in-paralog expression correlation differing from *B. plebeja*

in-paralog expression correlation by less than 0.1. Duplicate pairs from this group included both high and low expression correlation in both species; where in-paralog expression in *B. conchifolia* was conserved, the same was true of *B. plebeja*, and likewise for in-paralogs which had low in-paralog expression conservation. While a substantial proportion of the gene pairs were similarly conserved in expression between *B. conchifolia* and *B. plebeja*, there were a number of outliers. Outlying genes were biased towards one species in the conservation of expression between in-paralogs; some outlying duplicate pairs had highly correlated expression in *B. plebeja* in-paralogs but were more divergent in *B. conchifolia* in-paralog expression, and vice versa.

### 3.3.3 Annotation expressionally divergent loci

Gene duplicate pairs where in-paralogs have retained similar expression in one species but have diverged in expression in the other species may represent duplicated genes which were involved in adaptive divergence of *B. conchifolia* and *B. plebeja*, gene duplicates which fit this pattern were investigated further. Gene pairs with a difference of more than 0.7 between *B. conchifolia* in-paralog expression correlation and *B. plebeja* in-paralog expression correlation were extracted and functionally annotated by transferring annotations from the *B. conchifolia* genome (Table 1). 19 duplicate pairs, originating from 17 gene families were identified as having differentially conserved expression patterns between *B. conchifolia* and *B. plebeja*. Four of the gene families encoded genes which were involved in response to stress; Beta Glucosidase 17, Glutathione S-transferase TAU 19, Chalcone Synthase, and Tetraspanin family protein. Three gene pairs, representing two gene families, encode genes functioning in cell architecture and cell division (Tubulin Beta-5 Chain, Tubulin Beta Chain 4, and Actin-11).

| Gene Pair Family Size | OG | *B. conchifolia* Correlation | *B. plebeja* Correlation | Function |
|---|---|---|---|---|
| 11 | OG0000042 | 0.1090754 | 0.8344947 | tubulin beta-5 chain |
| 11 | OG0000042 | 0.545766 | -0.3848826 | tubulin beta chain 4 |
| 8 | OG0000052 | 0.9365615 | 0.1151401 | Pectin lyase-like superfamily protein |
| 4 | OG0000071 | 0.289671 | -0.5915618 | beta glucosidase 17 |
| 7 | OG0000121 | 0.2125419 | -0.5891892 | actin-11 |
| 7 | OG0000155 | 0.07904751 | 0.9755411 | 3-ketoacyl-CoA synthase 11 |
| 7 | OG0000155 | 0.1580627 | 0.9628352 | 3-ketoacyl-CoA synthase 1 |
| 5 | OG0000188 | -0.2011308 | 0.7587181 | delta 9 desaturase 1 |
| 7 | OG0000225 | 0.08268594 | 0.8214151 | glutathione S-transferase TAU 19 |
| 4 | OG0000571 | -0.1835204 | 0.9211444 | UDP-Glycosyltransferase superfamily protein |
| 2 | OG0000706 | -0.2943743 | 0.7880476 | Subtilisin-like protease |
| 6 | OG0000724 | 0.9553908 | -0.2545802 | PATATIN-like protein 4 |
| 3 | OG0000747 | -0.2544221 | 0.686826 | Phospho-2-dehydro-3-deoxyheptonate aldolase 1, chloroplastic |
| 3 | OG0001201 | -0.3554706 | 0.8507483 | Pyrophosphate-energized vacuolar membrane proton pump |
| 3 | OG0001764 | 0.6541896 | -0.08394763 | Chalcone synthase 1 |
| 4 | OG0001783 | 0.4364724 | -0.3186905 | unknown protein |
| 2 | OG0001951 | 0.7034487 | -0.5049693 | Tetraspanin family protein |
| 2 | OG0002541 | 0.2346892 | 0.9420487 | B-box zinc finger family protein |
| 2 | OG0003062 | -0.250498 | 0.6128935 | Plant invertase/pectin methylesterase inhibitor superfamily |

Table 1. Annotations and in-paralog expression Pearson correlation coefficient of *B. conchifolia* or *B. plebeja* in-paralogs found to have divergent expression correlation. Divergent in-paralog pairs were defined as divergent if they had a difference in correlation greater than 0.7. OG denotes the orthologous group the locus pair belong to.

### 3.3.4 GO term enrichment

To test whether any of the genes identified as being divergent in expression correlation were enriched for any functional terms, a GO enrichment analysis was conducted.

*C. sativus,* as the closest sequenced relative of *Begonia* available on the AgriGO server, was used as a background reference. The taxonomic proximity and therefore the most similar duplication patterns made *C. sativus* a suitable reference to use for *Begonia*.

A single closest homolog for each gene pair to be analysed in *C. sativus* was identified using BLASTX. The nucletotide sequences for selected *B. conchifolia* and *B. plebeja* sequences were used to query the *C. sativus* primary transcript peptide sequences with no E value threshold. *C. sativus* hits were retained if the match percent identity exceeded 70% and the length of the match was over 600 base pairs.

Of the 19 *Begonia* duplicate pairs chosen for enrichment analysis, 11 had an identified hit in *C. sativus*. The sequence names of the resulting 11 *C. sativus* hits were used as input to AgriGO for GO enrichment analysis.

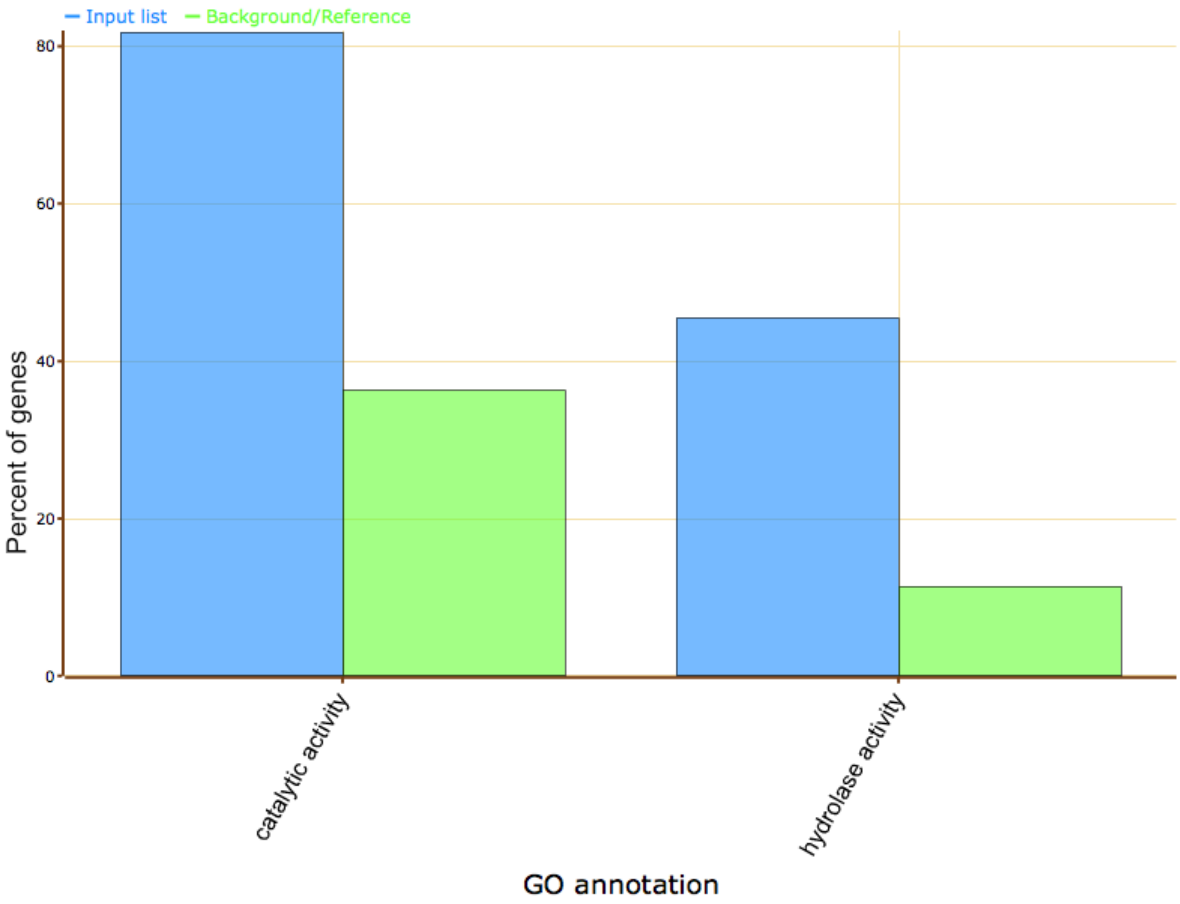



Figure 9. Percent of genes annotated with significantly over-represented GO categories in input gene list and in background gene list.

Catalytic activity and hydrolase activity were significantly overrepresented in the input set of *C. sativus* sequences representing *Begonia* duplicate pairs which have divergent in-paralog expression correlation between *B. conchifolia* and *B. plebeja* (alpha = 0.05, correction for multiple testing = FDR). To get a more detailed picture of the genes involved in the GO terms that are enriched, the *C. sativus* genes underlying the GO terms were extracted (Table 2).

| *C. sativus* sequence | Function |
|---|---|
| Cucsa.032090.1 | Pectinesterase/Pectinesterase inhibitor 34-related |
| Cucsa.102940.1 | Proprotein convertase subtilisin/Kexin |
| Cucsa.155000.1 | Pyrophosphate phosphohydrolase |
| Cucsa.155940.1 | Naringenin-chalcone synthase/Flavonone synthase |
| Cucsa.160130.1 | Phospho-2-oxo-3-deoxyheptonate aldolase |
| Cucsa.166940.1 | Purine-nucleoside phosphorylase |
| Cucsa.254780.1 | 3-Ketoacyl-CoA Synthase 17-Related |
| Cucsa.259630.1 | Very-long-chain beta-ketoacyl-CoA synthase |
| Cucsa.303140.1 | Tubulin |

Table 2. Annotations of *B. conchifolia* in-paralogs that were found to be divergent in either *B. conchifolia* or *B. plebeja* and were annotated with an over-represented GO term. Divergent in-paralog pairs were defined as divergent if they had a difference in correlation greater than 0.7.

Two gene families were chosen as candidates based on putative relevance to ecological fitness; 3-Ketoacyl-CoA synthase and Naringenin-Chalcone Synthase.

3-Ketoacyl-CoA synthase is a component of cuticular wax and suberin synthesis, making part of the fatty acid elongase complex which catalyses the biosynthesis of cuticular wax. Cuticular wax is an important for drought tolerance, and drought stress signals can act to upregulate genes involved in wax synthesis (Seo and Park, 2011). With a large diverse gene family in *A. thaliana*, mutants for some 3-Ketoacyl-CoA Synthases have shown reduced low humidity stress, though evidence suggests that there is redundancy between gene family members (Blacklock and Jaworski, 2006). Naringenin-Chalcone Synthase is a key enzyme in the anthocyanin biosynthesis pathway, catalysing the formation of naringenin chalcone from 4-coumaroyl CoA and malonyl CoA. The anthocyanin biosynthesis pathway is important for the production of pigments used for various stress related functions such as photoprotection and herbivore defence.

### 3.3.5 Analysis of the Chalcone Synthase and 3-Ketoacyl CoA Synthase families


Figure 10. Gene tree of 3-Ketoacyl CoA Synthase sequences identified in *B. conchifolia*, *C. sativus, A. thaliana* and *G. max.* MrBayes was used to reconstruct the phylogeny using a matrix of 1580bp. *Amborella trichopoda* 3-Ketoacyl CoA Synthase is used as an outgroup.

A phylogeny of 3-Ketoacyl-CoA Synthase (Figure 10) reconstructed using Bayesian inference identifies three major clades; one early branching clade and two derived from later duplications. All three clades include members of both *Begonia* and outgroup 3-Ketoacyl-CoA Synthase. Seven duplications giving rise to *Begonia* wax gene members occurred before the divergence of *Begonia* and the outgroup species, while nine duplications were specific to *Begonia*. Of the nine *Begonia* specific duplications, five appear as a duplication in *Begonia* giving rise to two duplicates next to a single unduplicated ortholog in *C. sativus*. Of the 34 *G. max* 3-Ketoacyl-CoA Synthase duplicates, twelve occur in a *G. max* specific duplicated cluster of four, supporting previous findings of a tetraploidy event in the recent evolutionary history of the species. Similarly, twelve duplicates out of twenty two in *A. thaliana* are derived from species specific duplications, reflecting the duplication history of the species. Five duplicates of the fifteen *C. sativus* members are derived from duplications specific to *C. sativus*. The *C. sativus* specific duplications give rise to two groups of paralogs; one in a cluster of three and one as a pair, likely representing tandem duplications due to the lack of a recent genome duplication in *C. sativus*.

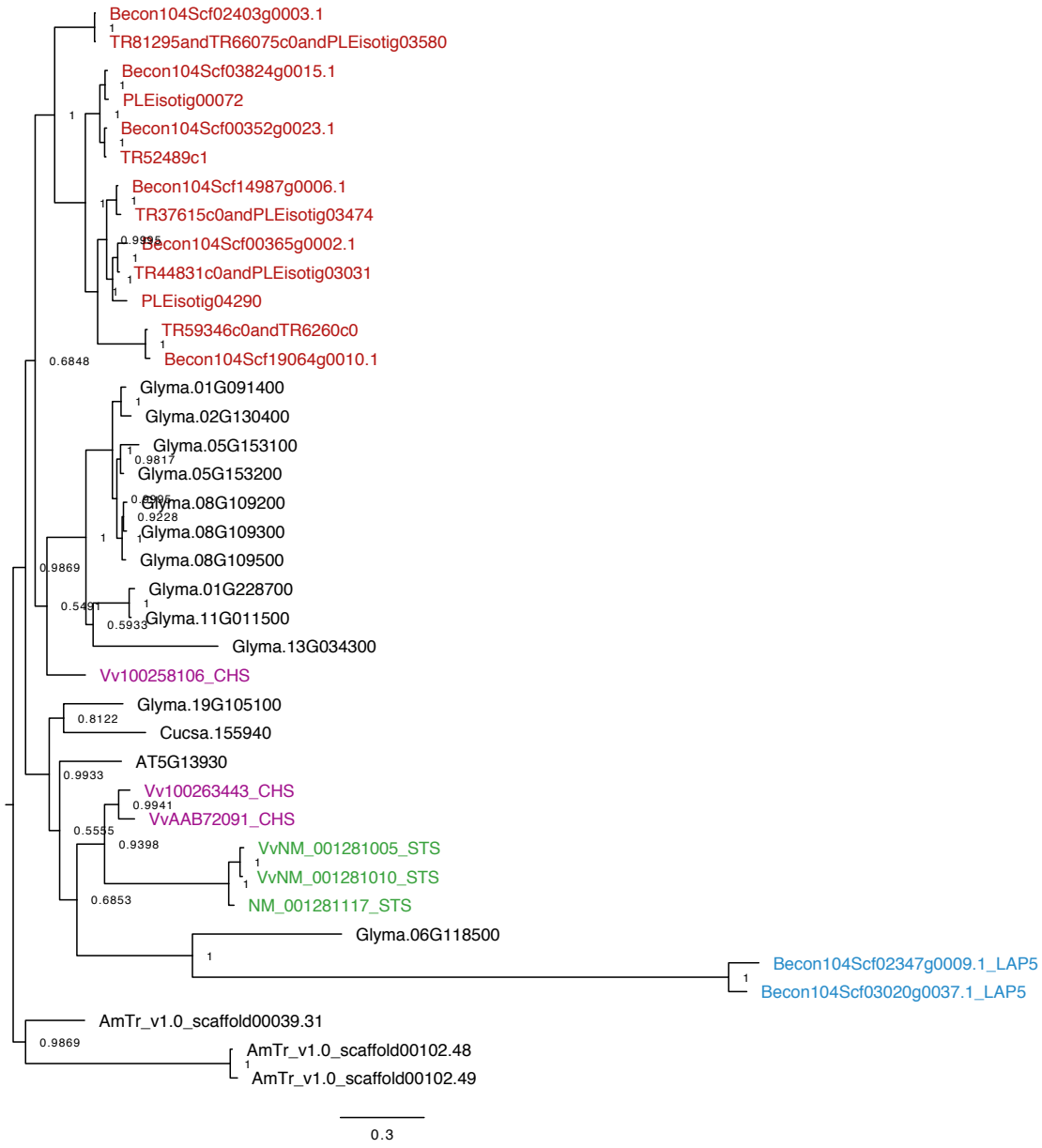



Figure 11. Gene tree of Chalcone Synthase sequences identified in *B. conchifolia*, *C. sativus, A. thaliana* and *G. max*. MrBayes was used to reconstruct the phylogeny using a matrix of 1304bp. Three copies of *Amborella trichopoda* Chalcone Synthase is used as an outgroup. CHS sequence from *Vitis vinifera* (pink), CHS-like Stilbene Synthase (STS) sequence from *V. vinifera* (green) and CHS-like LAP5 genes identified in *B. conchifolia* (blue) were included to confirm *Begonia* CHS orthology (see methods for details).

Reconstructing a Bayesian phylogeny of *Begonia* CHS (Figure 11) sequences with *A. thaliana*, *G. max* and *C. sativus* CHS sequences, using *Amborella* CHS as an outgroup showed the *Begonia* CHS gene family formed a monophyletic group, all duplicates having originated after the divergence of *Begonia* from its closest compared relative *C. sativus*. CHS-like LAP5 in *B. conchifolia* does not cluster with the identified *Begonia* CHS sequences, though they do cluster with the identified *A. thaliana* and *C. sativus* CHS sequences. This finding suggests that LAP5 and CHS diverged much less recently than CHS and STS in *V. vinifera*, and that the sequence shares greater similarity with the common ancestor of *Begonia* and *Vitis* CHS than with *Begonia* CHS. The retention of *Begonia* CHS monophyly, however, indicates that the sequences identified are indeed CHS and not a CHS-like sequence. Of the other comparison species, only *G. max* had multiple copies of CHS, not occurring as a monophyletic group; one duplicate being orthologous to the single copies of *A. thaliana* and *C. sativus* CHS, while the other copies formed an independent clade.

## 3.4 Discussion

### 3.4.1 Gene family reconstruction

This chapter set out to address the overwhelming ecological diversity seen across the genus *Begonia*, using two closely related but ecologically divergent species, *B. conchifolia* and *B. plebeja*. The study aimed to avoid *a priori* assumptions about the underlying environmental pressures that may be driving ecological diversification in *Begonia* in order to identify all environmental effectors important in *Begonia* evolution. This aim prompted a global, genome wide analysis, requiring the generation of both whole genome and transcriptome datasets. Using these resources, sequence divergence and differential expression data was mined from all *Begonia* gene families, with three outgroup species, *A. thaliana*, *G. max*, and *C. sativus* included for comparison. Using the OrthoFinder software, primary transcripts from *B. conchifolia*, *G. max*, *A. thaliana* and *C. sativus* were clustered into gene families. The OrthoFinder software was chosen because it minimises biases introduced by different taxonomic distances between input sequence species. Obtaining a set of gene families for *B. conchifolia* and outgroup species allowed a direct comparison of gene family sizes across species, and therefore the extent to which each species' genome is duplicated. The draft genome of *B. conchifolia* cannot be considered to be complete, and therefore data based on copy number counts must be interpreted with caution. While the majority of genes are sequenced, loci missing from the assembly and the predicted transcripts may artificially inflate or underestimate gene family sizes. Methodological problems could also have an effect on interpretation of results; while OrthoFinder successfully assigns gene family members to the correct group, it does not always identify all members of a gene family, possibly due to more divergent gene family members not meeting the thresholds set in the

clustering steps. Lack of sensitivity results in an artificially large number of single copy gene families, many of which belong to multigene families but have not successfully been clustered.

### 3.4.2 Comparative duplication patterns

*B. conchifolia* showed very little evidence of widespread gene duplication, a pattern mimicked by it's closest outgroup relative *C. sativus*. The genome of *C. sativus* has been noted in previous studies to lack a recent genome duplication, and therefore has few duplicate genes. Contrary to this pattern in *C. sativus, Begonia* has been shown to have a highly labile genome; high variation is seen in both chromosome number and genome size both within and between sections of *Begonia* (Figure 1), suggesting elevated rates of chromosome fission and fusion and large scale duplication events. Orthologous gene pairs from *B. conchifolia* and *B. plebeja* show that the two species, as well as a more distantly related South East Asian species *B. venusta,* share a genome duplication (Brennan et al. 2012). Evidence collected so far, therefore, suggests that unlike *C. sativus*, *B. conchifolia* has experienced duplication in it's evolutionary history, and that duplications are a prominent feature of the *Begonia* genus in general. One explanation for the apparent paucity of gene duplications in *Begonia* may be the result of a mixture of recent and recurrent duplication events, and poor assembly strategies, resulting in a large incidence of paralog collapsing. Such underestimation of duplication events may be a wider phenomenon, especially amongst crop species, with a high frequency of polyploidy (Renny-Byfield and Wendel 2014). *A. thaliana* appears to have more duplicated gene families than *B. conchifolia* and *C. sativus.* This is not unexpected, as *A. thaliana* has been shown to have undergone a number of genome duplications (Simillion et al. 2002). Despite the elevated level of duplicated gene families, *A. thaliana* has a much

smaller genome size than *Begonia*, suggesting some of the large fluctuations in genome size in *Begonia* could be due to transposon activity rather than genome or segmental duplications, a hypothesis which has preliminary support based on pilot studies (Arnau-Soler and Kidner, unpublished).

Gene family size was positively associated with pairwise divergence within gene families, though this trend was only apparent in gene families of five members or fewer; median pairwise sequence divergence levelled off with gene family size increases past this family copy number, and the same trend was observed in *C. sativus, A. thaliana* and *G. max.* Theoretical models have predicted that gene duplication decreases selection pressure on both duplicates, therefore allowing an increased tolerance to mutations (Kondrashov et al. 2002). However, the retention of sequence homology among even distantly related orthologs suggests that this tolerance is not finite. The limit of sequence divergence increase past a gene family size of 5 may reflect a mutational constraint. The trend seen in *Begonia* was mirrored in all the outgroup species, suggesting that similar processes govern the upper limits of mutation rate in coding sequences. For example, with greater numbers of mutations, the probability that a mutation causes a loss of function or a premature stop codon increases.

It is not surprising to note that, as sequence evolution appears to co-segregate with gene family size within limits, so does expression variation.

### 3.4.3 Ecological relevance of candidate genes

After functionally annotating the duplicate pairs which were divergently conserved between *B. conchifolia* and *B. plebeja*, an enrichment analysis identified which genes were significantly overrepresented. The genes which were of greatest interest in the context of

adaptive value were 3-Ketoacyl CoA-Synthase and Chalcone Synthase. 3-Ketoacyl CoA-Synthase is part of a family of membrane bound very long chain fatty acid (VLCFA) elongase enzymes. Anchored by two transmembrane domains (Ghanevati et al. 2001), 3-Ketoacyl CoA-Synthase elongates acyl substrates to make longer more complex VLCFAs via decarboxylation of malonyl-CoA groups. *A. thaliana* has a large gene family of 3-Ketoacyl CoA-Synthases, and the family members show varying amounts of substrate specificity towards longer or shorter chain fatty acids, as well as varying preference for saturated and unsaturated fatty acids. (Blacklock and Jaworski, 2006). Interestingly, the homologs of plant 3-Ketoacyl CoA-Synthases in yeast, belonging to the *Elop* gene family, and also being responsible for the synthesis of long chain fatty acids, appear to have highly similar mechanisms of catalysis. *Saccharomyces cerevisiae* mutants lacking two *Elop* genes required for the generation of fatty acids were rescued by the heterologous expression of *A. thaliana* FAE-1-like 3-Ketoacyl CoA-Synthase (Paul et al. 2006). Wax in plants (as well as in non-plant species (Gefen et al. 2015)) is an important component of the cuticle and epidermal coating of aerial tissues, preventing water from escaping the plant via non-stomatal means. Experimental data from *A. thaliana* has shown that a mutant for genes involved in the production of VLCFAs made young seedlings more sensitive to drought stress compared to wild type (Todd et al. 1999). Studies in *M. truncatula* and *A. thaliana* have also shown wax deposition to be a dynamic response to drought stress in plants, many transcription factors such as *WXP1* and MYB96 identified in each species respectively to be responsible for increased expression of genes associated with VLCFA synthesis (Zhang et al. 2005, Seo et al. 2011). Chalcone synthase, a homodimeric enzyme responsible for the first committed step in the synthesis of anthocyanins, is integral to the formation of myriad products responsible for plant pigmentation, UV protection and signalling. Using 4-coumaroyl CoA as a substrate molecule, CHS catalyzes three sequential condensation reactions of acetate from malonyl

CoA to produce naringenin-chalcone. Anthocyanins have been shown to have an antioxidant role, ameliorating reactive oxygen species (ROS) produced during plant stress responses (Manetas et al. 2002). Coloured pigments in particular have been shown to have an effect on ROS scavenging, red and green parts of *Pseudowintera colorata* leaves showed red leaves had lower levels of stress induced H2O2 and showed faster depletion of $H_2O_2$ after wounding compared to green parts of leaves (Gould, McKelvie and Markham, 2002).

Anthocyanins used for predator evasion has also been suggested, the green-absorbing effect of anthocyanins could act to prevent leaves reflecting light within the visible range of herbivorous insects (Manetas, 2006). Other studies, however, suggest that insects can discriminate between anthocyanic leaves and green leaves. The pleiotropy effect (Schaefer and Rolshausen, 2006, Lev-Yadun and Gould, 2009), where insects learn to avoid red leaves due to association with toxicity, and the handicap signal hypothesis (Archetti and Brown, 2004), proposing plants produce red leaves as an honest signal to insects of their defensive capabilities, have some support in senescing leaves (Karageorgou et al. 2008, Archetti, 2009), though this pattern is not seen in young red leaves, indicating distinct mechanisms driving anthocyanin production in this case (Kachi et al. 2004, Karageorgou et al. 2008).

The expression of the 12 3-Ketoacyl CoA Synthase genes with available expression data is highly diverse in the six tissues studied, with no one gene being dominant for all tissues. This is in stark contrast to Chalcone Synthase, where one copy (BeconScf00352g0023.1) is expressed at a very high level across all tissues in both species. With the exception of the male flower, CHS Becon00352g0023.1 is expressed in a highly similar pattern across all tissues, having an almost twofold increase in expression in *B. plebeja* male flower compared to *B. conchifolia* male flower. The remaining genes are more divergent in expression patterns between the two species, showing evidence of incipient reciprocal silencing in *B. plebeja* in female flower and leaf, and in root and vegetative bud. More expressional variation in CHS

copies that are expressed at lower levels dovetails with previous work finding that highly expressed genes evolve slower, possibly due to selection for higher translational robustness (Drummond et al. 2005, Pál et al. 2001). It is interesting to speculate that, while a main CHS locus fulfils most of each tissues requirements, the remaining loci may be functionally diversifying while being relatively unexposed to selection.

The 3-Ketoacyl-CoA Synthase family expresses more duplicate loci than CHS, with at least four loci with strong (> 300) relative expression in at least one tissue. Male flower has the largest relative expression of any one gene, indicating this tissue is extremely transcriptionally active across a range of genes. In *B. conchifolia*, two paralogs derived from a duplication specific to *Begonia* are the main loci expressed in male flower, whereas in *B. plebeja*, this pair of loci as well as another paralog pair, also derived from a *Begonia* specific duplication, are highly active in this tissue. The increased transcript abundance and greater transcript diversity in *B. plebeja* male flower reflects a change that has happened relatively recently, since the divergence of *Begonia* from the outgroup species. The retention of expression in *B. plebeja* of the four paralogs from both duplications, yet the retention of only two at such high levels in *B. conchifolia* suggests selection for greater transcript abundances in the former but not in the latter. The much larger flower size in *B. plebeja* may play a part in the demand for more 3-Ketoacyl-CoA Synthase activity. Root in both *B. conchifolia* and *B. plebeja* has one main locus expressed, which is likely a 3-Ketoacyl-CoA Synthase responsible for suberin production (Lee et al. 2009), being expressed at relatively low levels across all other tissues. While most 3-Ketoacyl-CoA Synthases in *B. plebeja* leaf are expressed at modest levels, one locus derived from a *Begonia* specific duplication is highly expressed in *B. conchifolia*. The single locus being responsible for the majority of 3-Ketoacyl-CoA Synthase transcripts in *B. conchifolia* leaves suggests this duplicate may

underlie *B. conchifolia*'s waxy leaves, though validation is required before this can be said conclusively.

Difference in expression diversity is accompanied by a very different pattern of duplication; all Chalcone Synthase duplications occurred after the divergence of *Begonia* from the outgroup species, appearing as a burst of duplication in both *Begonia* species. The 3 Ketoacyl CoA Synthase phylogeny on the other hand contains a large number of deep duplications predating the divergence of *Begonia* from all outgroup species, as well as duplications specific to *Begonia*. Effects of duplication age could underlie the difference in expression diversity seen between 3-Ketoacyl-CoA Synthase and CHS. A longer period of time since the birth of older 3-Ketoacyl-CoA Synthase duplicates provides more opportunity for duplicate genes to become co-opted into regulatory networks, whereas more recent duplications may be prone to an initial period of silencing.

The identification of two gene families, encoding genes responsible for anthocyanin biosynthesis and for wax deposition, which have divergent patterns of expression between duplicate genes has revealed functions which may have facilitated the divergence of *B. conchifolia* and *B. plebeja* into different niches. The observation that *B. plebeja* is found in drier, more open habitats, and *B. conchifolia* in the rainforest understorey supports anthocyanin biosynthesis being a divergent trait between the two study species. Growing in less shaded areas, *B. plebeja* has concerns of efficient sunlight use without the concomitant deleterious effects of photobleaching, associated with prolonged sun exposure. The blotchy anthocyanin patterning on *B. plebeja* leaves may act to buffer the negative effects of high direct sunlight by preventing photobleaching and scavenging free radicals.

*B. conchifolia* has less need for such high levels of anthocyanins in leaves, and this is reflected in the leaf morphology; dark green waxy leaves have anthocyanic tissue only at the

meeting of the leaf and the petiole. An understorey habitat provides different photo-efficiency concerns for species such as *B. conchifolia*, such as making best use of high intensity bursts of light in the form of sunflecks. It is evident, therefore, that the two species have divergent selective pressures acting on the, suggesting this may be driving the duplication and expression patterns. Ecological pressures that may be driving the diversification of 3-Ketoacyl CoA Synthase expression may also be explained by the different habitats *B. conchifolia* and *B. plebeja* occupy. Though intuitively it seems that *B. plebeja*, living in a more drought prone environment, would benefit from greater wax deposition, this species has a much less waxy leaf than *B. conchifolia*. Previous studies have found, however, that a greater amount of wax deposited on a leaf does not correspond to a proportional decrease on non-stomatal water loss (Riederer and Schreiber, 2001). This may explain why *B. plebeja* does not have a waxier leaf than *B. conchifolia*. Rather than drought tolerance, the use of wax in *B. conchifolia* leaves may help water run-off, and in doing so, reduce the ability of fungal spores and other pathogens to adhere to the leaf surface, thereby reducing the incidence of harmful pathogen and fungal attack.

### 3.4.4 Further work

Studying duplication patterns in *Begonia* species at greater resolution will be important in understanding the rate at which new genes are created. Sampling CHS and 3-Ketoacyl-CoA Synthase gene copy number and expression profile at the population level can reveal whether duplications are polymorphic in copy number and the how they are distributed across the population's range. If the turnover of duplications in these genes and other genes of ecological interest is high, it will be interesting to know whether this has any discernible effects on species stress responses, and in turn, whether any within-population differentiation

is possible from the effect of gene duplicates. Such findings would lend greater support to the importance of duplicate genes in diversification of *Begonia*. Given the large diversity in C-value seen in *Begonia*, both across and within sections, it will be interesting to expand this to more species, and to investigate whether any species have a ploidy series. Recurrent formation of polyploids has been shown to create widespread chromosomal and genomic variation (Buggs et al. 2012), and detection of unreduced pollen in *Begonia* species suggests that genome duplication has the potential to be common, thus it is of great interest to determine the rate at which such large scale changes happen in *Begonia* populations. Additionally, the identification of a ploidy series would provide an opportunity to study the fates of genes in polyploids compared to their diploid progenitors.

Divergent expression patterns between *B. conchifolia* and *B. plebeja* have been found in both CHS and 3-Ketoacyl-CoA Synthase, however a more controlled experiment is required to understand how the duplicate genes are differentially expressed between the two species. Using qRT-PCR, the response of *B. conchifolia* and *B. plebeja* to light and drought stress can be dissected across condition ranges and tissues. Results from such a study will be able to confirm the evolution of divergent mechanisms of stress response at duplicate loci.

# Chapter 4: Investigating the Chalcone Synthase gene family in *Begonia*

## 4.1 Introduction

### 4.1.1 Role of anthocyanins in plants

Plant pigmentation can often be striking, producing brightly coloured flowers and variegated leaves. One of the main compounds underlying these patterns is the group called the anthocyanins. Anthocyanins belong to the larger group of compounds called the flavonoids, plant phenolics with significant antioxidant properties (Heim et al. 2002), and there are 17 naturally occurring anthocyanins known in plants which are determined by the combination of aromatic or aliphatic sides chain on the sugars within the compound (Kong et al. 2003). Anthocyanins have been shown to have an antioxidant role, ameliorating reactive oxygen species (ROS) produced during plant stress responses. The B ring of the three benzene rings forming the backbone of flavonoids is one of the greatest determinants of the free radical scavenging activity; it donates hydrogen and an electron to ROS and RNS (reactive nitrogen species), thereby stabilizing them (Cao et al. 1997).

The first committed enzyme in the anthocyanin biosynthetic pathway (ABP) (Figure 1), Chalcone Synthase, has been shown to be upregulated by both local and systemic damage to white spruce, confirming the increased production of anthocyanins in response to stress (Richard et al. 2000). Photoprotection is another aspect of anthocyanin's protective roles. Anthocyanins often accumulate in photosynthetic tissues, allowing PSII to maintain function at higher levels of radiation stress (Smillie and Hetherington, 1999). Distinct to ROS scavenging, anthocyanins provide photoprotection by modulating the amount of different wavelengths of light reaching chlorophyll molecules, absorbing green, blue and UV light. The effect of this is the prevention of photoinhibition, which is

the overexcitation of the photosynthetic machinery by too fast a rate of energy capture relative to electron transport and dissipation (Long et al. 1994, Steyn et al. 2002). Manetas et al. (2002) found that young leaves of *Rosa* sp. and *Ricinus communis* L. containing high levels of anthocyanins suffered less loss of PSII photochemical efficiency than mature leaves which had low anthocyanin content. Coloured pigments in general have been shown to have an effect on ROS scavenging; red and green parts of *Pseudowintera colorata* leaves showed red leaves had lower levels of stress induced $H_2O_2$ and showed faster depletion of $H_2O_2$ after wounding compared to green parts of leaves (Gould, McKelvie and Markham, 2002). Studies in other taxa support the role of coloured pigments including anthocyanins in ROS scavenging; identification and quantification of phenolics in red and green leaved morphs of *Elatostema rugosum* showed a much higher level of anthocyanins in red leaves than green. Separation of phenolics showed that anthocyanins exceeded the contribution of other antioxidants to total phenolics, the strong correlation between anthocyanin content and antioxidant activity both indicate antioxidants play a key role in the scavenging of ROS (Neill et al. 2002). Contrary to these findings, Esteban et al. (2008) used *Erytronium dens-canis* L. (which has red patches on leaves) and *Pulmonaria officinalisi* L. (with light green spots on leaves) to test whether leaf variegation patterns had a photoprotective role, finding that anthocyanin containing sections of leaves were more sensitive to photoinhibition than green sections. Though the study did not support the original hypothesis, the maintenance of anthocyanins in leaves in spite of deleterious effects on photoinhibition provide strong evidence of an adaptive role. Anthocyanins may also have a role in predator evasion. The green-absorbing effect of anthocyanins could act to prevent leaves reflecting light within the visible range of herbivorous insects (Manetas, 2006). Other studies, however, suggest that insects can discriminate between anthocyanic leaves and green leaves. The pleiotropy effect (Schaefer and Rolshausen, 2006, Lev-Yadun and Gould, 2008), where insects learn to avoid red leaves due to association with toxicity, and the handicap signal hypothesis (Archetti and

Brown, 2004), proposing plants produce red leaves as an honest signal to insects of their defensive capabilities, have some support in explaining the occurance of red pigments in senescing leaves (Karageorgou et al. 2008, Archetti, 2009), though this pattern is not seen in young red leaves, (Kachi et al. 2004, Karageorgou et al. 2008). The red pigmentation may also be a signal not to the herbivores but to their predators. Anthocyanins in the leaf may undermine the herbivore's camouflage exposing it to its predators (Lev-Yadun et al. 2004). The literature suggests the use of anthocyanins differs across taxa, and the varying needs imposed by biotic and abiotic factors is accommodated by the versatile utility of these compounds. This may explain the often contradictory evidence gathered from different species, where the use of anthocyanins in one species may not mirror their use in another.

### 4.1.2 Putative roles of anthocyanins in *Begonia* diversity

The genus *Begonia*, whilst being incredibly speciose (Chapter 1), is also very morphologically diverse. It is found in montane habitats throughout the tropics, with the exception of Australia. Species of the genus have diversified in their leaf morphology, growth form and tolerance to environmental variables. Leaf pigmentation patterns are one of the most strikingly diverse characters, making this genus a good model for studying the evolution of pigmentation. Understanding the underlying ecological drivers of the different pigmentation patterns could therefore help unravel the complex role of anthocyanins in plants.

*Begonia* leaves have pigmentation patterns that vary widely across the genus, including reddening along leaf margins, adaxial anthocyanic blotches, anthocyanin deposition in leaf hairs, and redness along leaf veins. A common pattern in *Begonia* as well as many other tropical understorey dwelling species is the deposition of pigment in the abaxial layer (Lee, Lowry and Stone, 1979). The benefits abaxial anthocyanin deposition may confer are not fully clear yet; in 2008 Hughes, Vogelmann and

Smith tested the 'backscatter' hypothesis, suggesting abaxial anthocyanins acted to maximise light absorption in understorey shade tolerant species by reflecting light which passes through the leaf back onto the mesophyll, thus maximising the light captured. The authors, however, found no evidence of this being the case. Rather than a backscattering role, Hughes et al. (2015) used differentially variegated leaves of *Colocasia esculenta* to show that abaxial anthocyanins may provide photoprotection of understorey plant leaves during periods of high irradiance without incurring costs of lower photosynthetic yield associated with adaxial anthocyanin deposition.

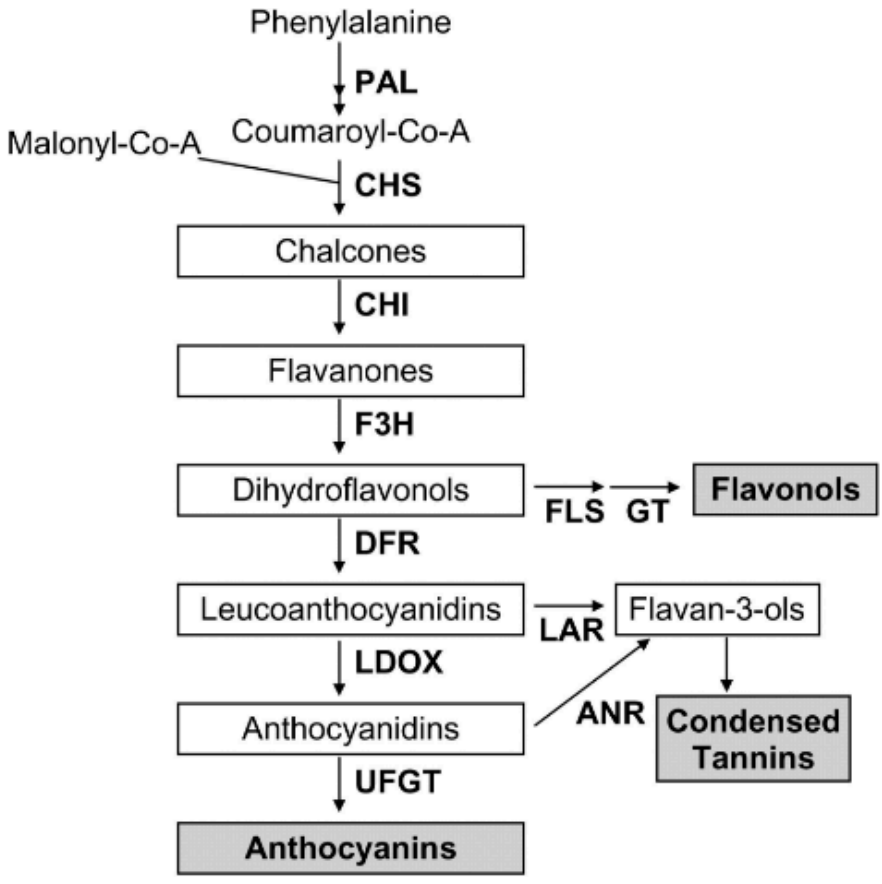



Figure 1. A schematic of the Anthocyanin Biosynthetic Pathway, adapted from Takos et al. 2006.

*Begonia* species are frequently found in deeply shaded habitats such as forests. Many are found on rocky limestone as creepers or weeds (e.g. *B. wuzhishanesis* and *B. siccaudata*), others are able to grow as epiphytes or terrestrially (e.g. *B. scottii*). Each of these species has anthocyanin pigmentation; in the leaf *B. scottii* has pigment in hairs on leaves and petioles, *B. siccaudata* has abaxial anthocyanin deposition, and *B. wuzhishanensis* has reddened abaxial and adaxial veins, as well as showing some evidence of iridescence. Some species of *Begonia* are found in more open areas, such as *B. samhaensis* and *B. socotra*, which live on shaded cracks on open limestone, though sometimes are also found terrestirially under cover of montane shrubland. Additionally, both these species have no anthocyanin pigmentation adaxially or abaxially.

Some association is seen in *Begonia* between habitat and anthocyanin patterning, though no distinct patterns emerge. The rapid transition between different pigmentation patterns suggests this trait could be of high adaptive value. The governing of anthocyanin patterns could also, however, be the result of genetic drift, which would equally explain the highly variable leaf patterning across the genus. The highly structured populations and poor dispersal shown in *Begonia* would also support a hypothesis of drift being responsible for rapid change and fixation of new pigmentation patterns across species in the genus. Research in other plants has shown a wide range of mechanisms controlling variation in anthocyanin patterning, a large proportion of which is regulatory. Studies in the basis of anthocyanin accumulation variation in *A. thaliana* (Schulz et al. 2015) showed a large amount of variation in anthocyanin biosynthetic gene transcript abundance across 54 accessions, with the known transcription factors PAP1 and PAP2 being correlated with abundance of transcripts, though other transcription factors not known to regulate anthocyanin biosynthesis genes MYB12, MYB11 and MYB111 were also found to be correlated with anthocyanin biosynthetic gene transcript abundance. High variation in transcript abundance, however, was not mirrored in leaf anthocyanin contents, suggesting post transcriptional regulation is an important regulatory process in anthocyanin

production. Competition of the anthocyanin pathway for substrate with the flavonol biosynthesis pathway has been shown to provide another layer of regulatory complexity in *Mimulus* (Yuan et al. 2016). Divergent patterns of flower anthocyanin deposition are achieved between two sister species *M. lewisii* and *M. cardinalis* by differential expression of the transcription factor LAR1, which positively regulates flavanol synthase (FLS), a gene in the flavanol biosynthesis pathway. FLS competes with DFR, a gene in the anthocyanin biosynthesis pathway, for substrate. White patches on the corolla of *M. lewisii*, important in pollinator signalling, are due to the depletion of anthocyanic pigments in these areas as a result of the upregulation of FLS by LAR1, thus redirecting substrate into the production of colourless flavonols rather than coloured anthocyanins.

| Species | CHS copy number |
|---|---|
| *R. communis* | 4 |
| *C. sativus* | 3 |
| *P. persica* | 9 |
| *M. esculenta* | 11 |
| *M. domestica* | 16 |
| *M. trunculata* | 21 |
| *L. usitatissimum* | 13 |
| *G. max* | 19 |
| *P. trichocarpa* | 15 |
| *P. vulgaris* | 16 |
| *F. vesca* | 6 |

Table 1. Selected Fabid species and the number of copies of Chalcone Synthase found in each species. Copy numbers were obtained by BLASTX of *A. thaliana* CHS cDNA sequence (AT5G13930) against all Fabids species in Phytozome, using an E value threshold of $1e^{-40}$

### 4.1.3 Premise of the chapter

To begin to unravel the causes underlying the rapidly evolving patterns of pigmentation in Begonia, this chapter will examine the anthocyanin biosynthetic pathway, the pathway responsible for production of coloured compounds such as anthocyanins. The focus of this chapter will be Chalcone Synthase (CHS), the first committed step of the anthocyanin biosynthetic pathway (Figure 1), due to it's importance in the pathway, it's extensive characterization and the interesting patterns of evolution of this enzyme in other species. This chapter will follow three lines of investigation into CHS and its role in diversification of leaf colour and pattern. Firstly, analysis of copy number will seek to find whether there are divergent patterns of duplication in different lineages of Begonia. Cloning techniques will be used to identify CHS copy number across a range of species in the genus. Secondly, evolutionary consequences of gene duplication include acquisition of functions distinct from the function of parent genes. Testing the hypothesis that positive selection has played a role in the evolution of the CHS gene family in *Begonia*, selection analysis will be used to test branches of interest in the CHS gene phylogeny. Finally, new expression patterns, rather than new functions, can also create novel phenotypes, and searching for divergent expression profiles between species may provide evidence of regulatory changes underlying new pigment patterns. Differences in expression patterns of CHS between *B. conchifolia* and *B. plebeja*, two closely related and phenotypically distinct species, will be investigated to shed light on whether differential regulation of genes may underlie phenotypic divergence.

## 4.2 Methods

### 4.2.1 CHS survey

Degenerate primers were designed in order to amplify all copies of CHS from a range of South American species (forward primer: AYCCDGAYTWSTACTTTCGS, reverse primer: TTCYKYCKCATCTCRTCCA), yielding a product size of roughly 1,200 base pairs. To design primers, all CHS sequences identified in the *B. conchifolia* draft genome assembly were used (see below). Sequences obtained were aligned using MAFFT v6 (Katoh et al. 2002), alignment quality was ensured by manual inspection. The Primer3 (Rozen and Skaletsky, 2000) plugin in Geneious v6.0.6 (Kearse et al. 2012) was then used to design primers across all sequences, primer results were picked based on the longest product and the least number of degenerate bases. Primers were tested on all species used by PCR (see program below), ensuring that only one band was present for each PCR reaction. DNA from 5 South American species (*B. solananthera*, *B. nelumbiifolia*, *B. odorata*, *B. cubensis*, *B. foliosa*) and one South East Asian species (*B. puspitae*) was obtained from the DNA bank at Royal Botanic Gardens Edinburgh (EDNA). Species from South America were chosen based on being in the same radiation as *B. conchifolia* and *B. plebeja*, as well as availability of DNA. Next generation sequencing data available for a number of South East Asian species (*B. burbidgei* and *B. venusta*) prompted the inclusion of *B. puspitae* from South East Asia for comparison.
Degenerate primers were used to amplify all copies of CHS as follows: 1ul of DNA was combined in a total volume of 20ul with 2ul 10uM CHS degenerate F primer, 2ul 10uM CHS degenerate R primer, 2ul PCR buffer (Bioline: 160mM (NH4)2SO4, 670mM Tris-HCl (pH 8.8 at 25$^{o}$C), and stabilizer), 1ul $MgCl_2$, 2ul 10mM dNTPs, and 5 units Taq polymerase (Bioline). The PCR program had an initial denaturing step of 94$^{o}$ for 1 minute, then 30 cycles of 94$^{o}$ for 1 minute, 55.9$^{o}$ for 1

minute and 72$^{o}$ for 1 minute, followed by 72$^{o}$ for 5 minutes. The PCR reactions were carried out in a PTC-200 Peltier Thermal Cycler (MJ Research). Successfully amplified regions were identified by a single band of a length around 1,100 base pairs, verified by running through a 2% agarose gel. After cleaning with Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare), the PCR products were ligated into a pGEM-T vector (Promega). 100ul of a-Select Chemically Competent Cells (Bioline) were transformed by the addition of 5ul of the ligation reaction. To control for variable transformation efficiencies, 100ul and 900ul of the cell mixture was plated onto LB agar plates containing 10ul 100ug/mL Ampicillin, and plates were incubated overnight at 37$^{o}$C.

50 clones picked at random were screened with colony PCR, using the PCR program described above and CHS degenerate primers. Sequence variation was used to distinguish between different copies of CHS by digestion of PCR products with SmaI and NaeI (New England Biolabs), two clones exhibiting each digestion pattern were chosen, where possible, for sequencing to reduce the risk of missing any copies of CHS. Each clone was sequenced twice using forward and reverse primers. To obtain the sequence of the selected plasmid inserts, undigested PCR product was cleaned using ExoSap-IT (Affymetrix). The sequencing reaction was carried out using the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher) on a PTC-200 Peltier Thermal Cycler (MJ Research), then purified and sequenced using a Prism genetic analyser (Applid Biosystems). Forward and reverse sequence pairs, where available, were visualized, aligned and manually edited in Geneious v6.0.6 (Kearse et al., 2012), using PHRED scores for quality assessment.

Copy number of CHS in each species used in the survey was estimated by eliminating sequences which were identical, retaining the longest representative of each putative locus sequenced (table 1). Copy number estimation may be confounded by failure to amplify some copies, and duplicates being discounted due to the absence of divergent coding regions due to incomplete coverage during sequencing.

Manually checked and trimmed CHS nucleotide sequences were aligned using MACSE 1.02 (Ranwez et al. 2011), a frameshift aware aligner. Ungapped nucleotide sequences extracted from the alignment were translated and aligned with the *A. thaliana* CHS homolog (AT5G13930) protein, and the resulting alignment was manually checked for quality. The catalytically important sites identified by Ferrer et al. (1999) were mapped onto the peptide sequence of the Alfalfa CHS2 peptide sequence (Uniprot ID P30074-1), and the annotations were transferred to *Begonia* by aligning the annotated Alfalfa CHS2 sequence with cloning and NGS derived *Begonia* translated nucleotide sequences.

### 4.2.2 Identifying members of CHS in the Fabids

The *A. thaliana* single member of CHS (AT5G13930) peptide sequence was used to query the Fabid species using the Phytozome comparative plant genomics platform (Goodstein et al. 2012). BLASTX was used with an E value threshold of $1e^{-40}$ (Altschul et al. 1990), hits found were aligned and manually inspected for homology across the length of the coding sequence.

### 4.2.3 Identifying CHS loci in the B. conchifolia genome sequence

A draft *B. conchifolia* genome assembled by Aurelianao Bombarely using PacBio and Illumina data (Chapter 2) was considered to be the most complete resource available. The *A. thaliana* peptide sequence for CHS, AT5G13930 was used to query the *B. conchifolia* predicted protein sequences using BLASTP with an E value cutoff of $1e^{-10}$. The hits were verified by using them to query the TAIR10 (Berardini et al. 2015) nucleotide database with BLASTN, ensuring the top hit was CHS AT5G13930. An alignment of the hits was built using MAFFT (Katoh et al. 2002) and manually inspected for homology across the sequence.

### 4.2.4 Identifying transcriptomic CHS sequences

Because no genome is available, CHS sequences from *B. plebeja* were obtained from the reference transcriptome assembled from RNA-seq reads generated using the Illumina sequencing platform (Chapter 2). Sequences were identified by using TBLASTN to query the *B. plebeja* reference transcriptome using the *A. thaliana* protein homolog (AT5G13930) using an E value cutoff of $1e^{-10}$. Sequences were aligned using the MAFFT plugin in Geneious v6.0.6 (Kearse et al. 2012) with the *A. thaliana* transcript sequence. *B. plebeja* sequences which had low homology to all other sequences were used to query the TAIR10 database in TAIR, if the reciprocal hits did not include the *A. thaliana* CHS homolog (AT5G13930) the *B. plebeja* sequences were eliminated from further study. Finally, sequences which shared very high homology (100% sequence identity over at least 200 bp) were treated as suspected alleles, all but the longest representative alleles were removed from further analysis.

### 4.2.5 Phylogenetic analysis

A gene tree of CHS was reconstructed using transcriptome derived CHS sequences in *B. conchifolia* and *B. plebeja*. A gene phylogeny of CHS sequences from other species of *Begonia* (obtained from the CHS PCR survey, see Section 4.2.1) is not currently available due to the high frequency of incomplete coding sequence obtained for CHS members. Sequences in the resulting alignment often do not overlap with many other sequences, therefore it was deemed necessary to undertake more sequencing before performing phylogenetic analyses. Sequences were aligned using MAFFT (Katoh et al. 2002) and the alignment was manually inspected for quality. A 1,236 base pair length of sequence was extracted from the alignment, corresponding to well aligned sequence excluding the 5' and 3' UTRs.

Different substitution models were tested to identify the most suitable method to use when building the phylogenetic tree. SYM and K80 were selected as the best models by AIC and BIC criteria respectively using jModelTest 2.1.2 (Posada, 2008). Due to implementation of GTR models only in RAxML v8 (Stamatakis, 2014) and the absence of invariant sites and gamma distribution of rates in the models selected by jModelTest, the GTR model was used for all partitions in phylogenetic analyses.

RAxML was used to reconstruct a maximum likelihood tree using the GTRCAT model, and testing support for each node using 1000 bootstrap replicates. A Bayesian tree was also reconstructed using Mr Bayes 3.2 (Ronquist and Huelsenbeck, 2003) with the GTR model. The analysis was run using two chains under default settings for 10,000 generations, sampling every 10 steps. Additional repeats of 10,000 generations were added until the standard deviation of split frequencies was below 0.01. The first 250 samples were considered burnin and were discarded. The trees produced by maximum likelihood and Bayesian methods were compared to ensure the topologies were consistent between the two methods.

### 4.2.6 Selection analysis

A codon alignment was created for codeml using Pal2Nal v14 (Suyama, Torrents and Bork, 2006). A region of the CHS coding region was extracted from all transcriptome derived *B. conchifolia* and *B. plebeja* sequences within Geneious v6.0.6 to exclude any poorly aligned UTR regions. The sequences were translated into all 6 reading frames using dna2pep (Wernersson, 2006), and aligned with the *A. thaliana* CHS peptide sequence, enabling selection of the correct reading frame for each sequence. Peptide sequences from *Begonia* and *A. thaliana* were aligned using mafft with the --anysymbol flag, and the resulting alignment and the nucleotide sequences were provided to pal2nal

for codon alignment. The –nogap flag was used with pal2nal to remove all stop codons from the alignment automatically. A maximum likelihood tree was made using RAxML using the GTRCAT model with 1000 bootstrap replicates, using *A. thaliana* as an outgroup. The branches to be tested were marked in the newick tree file using a '#1' symbol. For each branch, two codeml control files were set up, one for the null hypothesis (H0), where there was one dN/dS for the whole tree, and one for the alternative hypothesis (H1), where dN/dS varies between the focal branch and the rest of the tree.

For H0, model was changed to 0 (one dN/dS), and an unmarked tree was provided. For H1, model was set to two (2 or more dN/dS values for branches).

### 4.2.7 Expression Analysis

Expression of CHS copies used RNA-seq reads generated from 6 different tissues from *B. conchifolia* and *B. plebeja* (Chapter 2). Predicted transcripts from the *B. conchifolia* genome were used as a reference. Examination of the predicted transcripts revealed some duplicate sequences which appeared to originate from the same locus, therefore the transcripts were filtered first to prevent underestimation of expression levels. The tr2aacds pipeline within the EvidentialGene package (Gilbert, 2013) was used for filtering, the method uses CDS and peptide translations of the input transcripts to assign them to a filter pass set, a set of putative isoforms, and a dropped set, excluded from further analyses. For the purposes of this analysis, the filter pass set and the putative isoform set were concatenated to minimise the loss of useable loci due to exclusion.

The read mapping software Salmon (Patro, Duggal and Kingsford, 2015) used the filtered predicted transcripts as a reference to map reads against. After indexing the transcripts each replicate from each tissue was mapped to the reference transcripts with Salmon using default parameters.

Between sample normalisation was done using TMM normalisation within the EdgeR package (Robinson, McCarthy and Smyth, 2010). After obtaining the matrix of normalised expression values, the rows corresponding to CHS copies were extracted, and the means of the replicates per species were calculated.

## 4.3 Results

### 4.3.1 Chalcone synthase characterization in other species

Chalcone synthase (CHS, EC 2.3.1.74), a homodimeric enzyme responsible for the first committed step in the synthesis of flavonoids, is integral to the formation of myriad products responsible for plant pigmentation, UV protection and signaling. Using 4-coumaroyl CoA as a substrate molecule, CHS catalyzes three sequential condensation reactions of acetate from malonyl CoA to produce naringenin-chalcone (Figure 1). Four strictly conserved residues are thought to underlie CHS catalytic function based on a copy of Alfalfa CHS: $Cys^{164}$, $His^{303}$, $Asn^{336}$ and $Phe^{215}$ (Jez and Noel, 2000). The structure and catalytic mechanism of CHS was elucidated by Ferrer et al. (1999) using a recombinant Alfalfa CHS2 peptide sequence. Each homodimer subunit being capable of catalysing the production of chalcone (Tropf et al. 1995). Subunits connect along a roughly straight surface, at one terminal of this surface Met137 of one monomer protrudes and is buried within a pocket in the second monomer. The starter substrate p-coumaroyl CoA is initially bound by the coumaroyl binding pocket, composed of Alfalfa residues Ser 133, Glu 192, Thr 194 and Ser 338, the coumaroyl binding pocket residues are presumed to underlie the preference for binding coumaroyl as a substrate. The CoA binding tunnel also contributes to initial substrate binding, comprising Alfalfa residues Cys164, Lys55, Arg58 and Lys62 as well as Ala308. The coumaroyl binding pocket residues are presumed to underlie the preference for binding coumaroyl as a substrate. The reaction proceeds with the

disassociation of the CoA enzyme, and subsequent binding and decarboxylation of the first malonyl CoA molecule allows the transfer of the coumaroyl group to form a diketide. Production of chalcone requiring one p-coumaroyl CoA and three malonyl CoA molecules the reaction is repeated twice to elongate the diketide to a tetraketide, Cys164, which is activated by His303, plays a key role in the production of the tertraketide intermediate and both are found to be conserved across many CHS-like enzymes including in bacteria. Following the formation of the tetraketide intermediate, the molecule undergoes cyclization in another region of the enzyme located in the cyclization pocket, which is separated from the coumaroyl binding pocket by Alfalfa residue Phe265, and consists of Alfalfa residues Thr 132, Met 137, Phe215, Ile 254, Gly 256, Phe265 and Pro375. The size of the cyclization pocket physically restricts the intermediate substrate to three additions of malonyl coA. The cyclization of the tetraketide facilitates the release and aromatization of the substrate to form chalcone. In addition to the main residues which compose the active site, a number of other residues contribute to the geometry of the enzyme, including Pro138, Gly163, Gly167, Leu214, Asp217, Gly262, Pro304, Gly305, Gly306, Gly335, Gly374, Pro375 and Gly376.

High structural conservation is seen in CHS, with most species having one conserved intron position *(Antirrhinum majus* being an exception). In addition to the conservation of the four residues reported to be involved in the catalytic machinery, a conserved motif across chalcone and stilbene (a closely related polyketide synthase) synthases has been identified (WGVLFGFGPGLT) (Durbin et al. 1995), suggesting regions of conservation important for function. Despite widespread structural conservation, there is evidence of diversification among CHS copies within genomes. CHS copy number is variable among species, many having CHS encoded by multigene family, such as in *Ipomoea* (Dao et al. 2011), which has 6 copies (Durbin et al. 2000), whereas other species such as Arabidopsis and *Antirrhinum* having only one (Dao et al. 2011).

Preservation of multiple copies may have resulted in functional divergence between copies of CHS, multi-copy species showing expansion or switching of substrate preference, with consequent novel

products (Christensen et al. 1998). Molecular evolutionary analyses using *Ipomoea* CHS copies suggested that one of the two divergent groups of CHS loci had experienced relaxed purifying selection, and identified amino acids which may have been under positive selection (Yang et al. 2004). Amino acid sites around the active site were among the residues that were different between the two divergent groups, and may have been involved in their divergence. Expression analysis on these groups revealed differential expression both in tissues and developmental stages, pointing to additional expressional and/or regulatory divergence.

### 4.3.2 Chalcone synthase copy number variation in *Begonia*

| Species | Section | Distribution | Chromosome Number | Copy Number |
|---|---|---|---|---|
| *B. solananthera* | Solananthera | SE Brazil | 56 | 7 |
| *B. foliosa* | Lepsia | NW Venezuela to Ecuador | 60 | 8 |
| *B. odorata* | Begonia | Lesser Antilles | 52 | 11 |
| *B. nelumbiifolia* | Gireoudia | Mexico to Colombia | 28 | 7 |
| *B. puspitae* | Reichenheimia | Sumatra | ? | 5 |
| *B. cubensis* | Begonia | Cuba | 52 | 10 |
| *B. conchifolia* | Gireoudia | Cost Rica to Panama | 28 | 7 |
| *B. plebeja* | Gireoudia | CS Mexico to C America | 28 | 6 |

Table 2. Table of selected *Begonia* species, their section, distribution, chromosome number and CHS copy number.

### 4.3.3 Chalcone synthase catalytically important sites in *Begonia*

Translation of Begonia CHS sequences allowed mapping of catalytically important sites using the Alfalfa CHS protein sequence as a guide (Ferrer et al. 1999). Sites investigated had roles both in

direct catalysis as well as the maintenance of enzyme geometry. Residue substitutions were evaluated for their effect in terms of residue property using an amino acid exchangeability metric (Yampolsky and Stolzfus, 2005). Three residues which composed the coumaroyl binding pocket had been subsituted in at least two species. 3 members of *B. cubensis*, 4 members of *B. odorata*, *B. foliosa* and *B. nelumbiifolia* and 2 members of *B. puspitae* CHS had an S —> T substitution at Serine 133. Threonine 194 had a T —> S substitution and a T —> A substitution in *B. plebeja* and *B. cubensis* respectively. The other threonine residue at position 197 had T —> M substitutions in 2 members of *B. cubensis* and 1 member of *B. odorata*, and a T —> A substitution in 1 member each of *B. cubensis* and *B. odorata*. Serine 338 and Glutamic acid 192, the remaining two residues involved in coumaroyl binding, showed no change in any of the species for which sequence was available. Amino acid exchangeability values for amino acid substitutions were high in the main for coumaroyl binding pocket residues, most scores higher than 300, only T —> M substitutions at threonine 197 scoring lower at 261.

Two residues forming the CoA binding tunnel both had substitutions; Alanine 308 had 2 A —> G substitutions and 1 A —> C substitution, both in *B. cubensis*, and 1 member of *B. solananthera* had an A —> E substitution. Lysine 62 has K —> R substitutions in 1 member of *B. cubensis*, *B. odorata* and *B. foliosa*, 2 members of *B. nelumbiifolia* and 4 members of *B. solananthera*. Similarly to amino acid substitutions in residues forming the coumaroyl binding pocket, all CoA binding residue substitutions were between amino acids with high exchangeability (> 300), suggesting a small effect conferred by the changes.

Of seven amino acids forming the cyclization pocket, only three have changes in any species at the amino acid level. *B. solananthera* alone has any substitution at methionine 137, 2 members with M —> T substitutions, and 8 members with M —> I substitutions. Isoleucine 254 shows 1 I —> L substitution in *B. odorata* and 2 I —> V substitutions in *B. puspitae*. Phenylalanine has 1 F —> L substitutions in *B. nelumbiifolia*, 2 in *B. puspitae*, 4 in *B. cubensis*, 5 in *B. odorata*, and 6 in *B.*

*foliosa*. *B. nelumbiifolia* also has one member with an F —> S substitution. Whilst substitution exchangeability at phenylalanine 265 (~330) and Isoleucine 254 (301 and 537 for *B. odorata* and *B. puspitae* respectively) were all high, methionine 137 subsitutions in *B. solananthera* were lower (M —> T = 152 and M —> I = 279). The lower exchangeability between M and T was reflected in fewer copies which had this substitution (2) compared to the more exchangeable M —> I (8).

| Subs. | Exch. | *odorata* | *nelumbiifolia* | *foliosa* | *cubensis* | *plebeja* | *conchifolia* | *solananthera* | *puspitae* |
|---|---|---|---|---|---|---|---|---|---|
| S133T | 13 | 4 | 4 | 4 | 3 | 0 | 0 | 0 | 2 |
| T194S | 56 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| T194A | 60 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| T197M | 5 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| T197A | 60 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| A308G | 80 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| A308C | 52 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| A308E | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| K62R | 46 | 1 | 2 | 1 | 1 | 0 | 0 | 4 | 0 |
| M137T | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| M137I | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 |
| I254L | 60 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I254V | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| F265L | 24 | 5 | 1 | 6 | 4 | 0 | 0 | 0 | 2 |
| F265S | 21 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G262W | 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G262R | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| D217W | 4 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| D217Y | 49 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G305C | 57 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Table 3. Summary of substitutions found in *Begonia* CHS sequences used in PCR survey

No change was detected at residues threonine 132, phenylalanine 215 and glycine 256 in sequence data available. Conserved residues Cys164, Phe215, His303, Asn336 were also found to be conserved in all sequence data available in *Begonia.*13 residues were identified by Ferrer et al (1999) as being important for maintaining enzyme geometry, three of these showed change in at least one species, with the remaining ten being conserved across all species where the sequence data was available. Glycine 262 had 2 members of *B. odorata* showing G —> W substitutions, and 1 member

of *B. solananthera* with G —> R substitutions. 1 member of *B. solananthera* had a G —> C substitution at Gly 305. Asparagine 217 had substitutions in 2 members of *B. cubensis* from D —> W, and 1 member of *B. odorata* with a D —> Y substitution. Interestingly, while substitutions at glycine 262 and glycine 305 have reasonably high exchangeability values (~250), the D —> Y substitution in *B. odorata* has an exchangeability value of 87, and the D —> W substitution in *B. cubensis* lacks observational data to support statistical support of exchangeability. This finding suggests these substitutions may have a marked effect on enzyme geometry, and possibly function. Results are summarized in Table 3.

### 4.3.4 *Begonia* Chalcone synthase phylogenetics

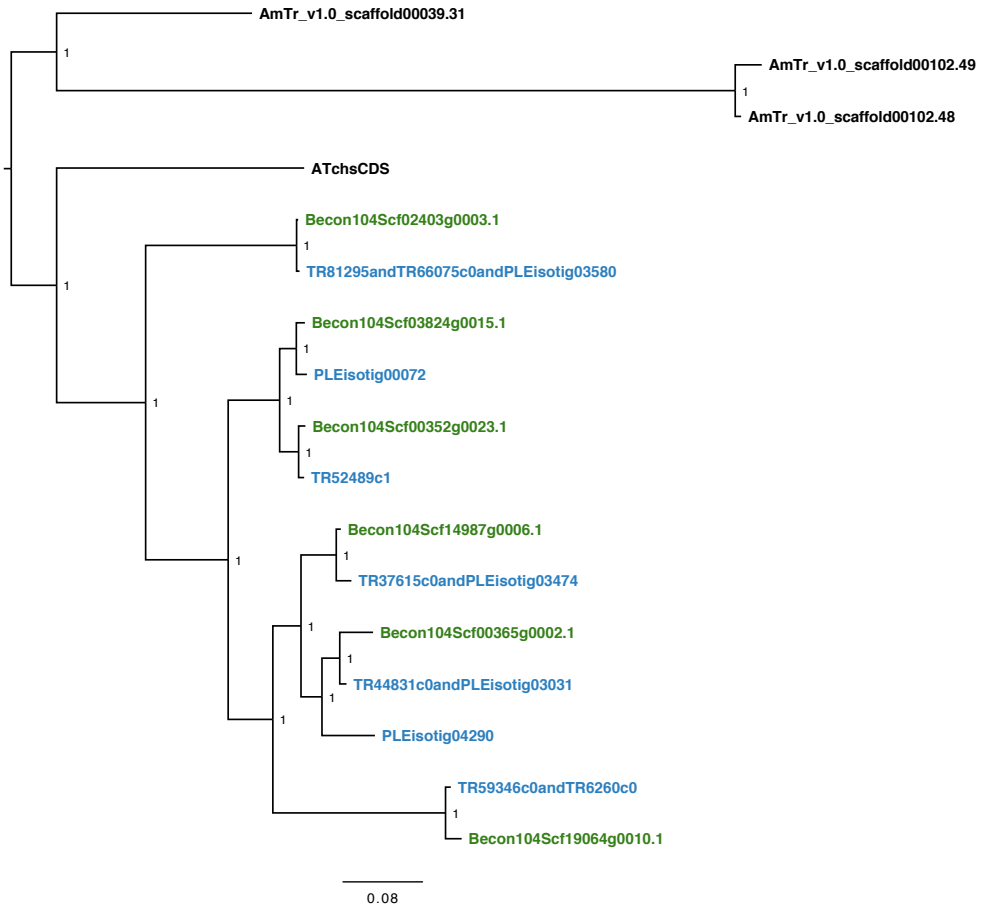



Figure 2. Gene tree of CHS sequences identified in *B. conchifolia* and *B. plebeja*. Reconstructed using MrBayes using a 1,236 bp matrix, using all three copies of *Amborella* CHS as an outgroup. *B. conchifolia* loci are coloured in green, *B. plebeja* loci in blue.

The CHS gene tree reveals the relationships of the paralogs in *B. conchifolia*, which underwent 5 duplications, and *B. plebeja*, with 6 duplications, showing that all duplications excluding the *B. plebeja* specific one occurred before the divergence of the two species. The first duplication resulted in two branches, one of which went on to undergo further duplications, while the other remained unduplicated and conserved between the two species, made evident by the short branch lengths in each ortholog. The branch leading to further duplications gave rise to two main clades. One clade consists of two duplicates each in *B. conchifolia* and *B. plebeja*, with comparative levels of divergence between species and between duplicates. The second clade has a more complex duplication history. A first duplication leads to two groups: one consisting of a copy in both species which appears to undergo rapid divergence after the duplication, evidenced by a long branch length. The other group produced contains one duplication which yields two copies in *B. conchifolia* and *B. plebeja,* and another duplication specific to B. *plebeja.*

4.3.5 Selection in *Begonia* chalcone synthase

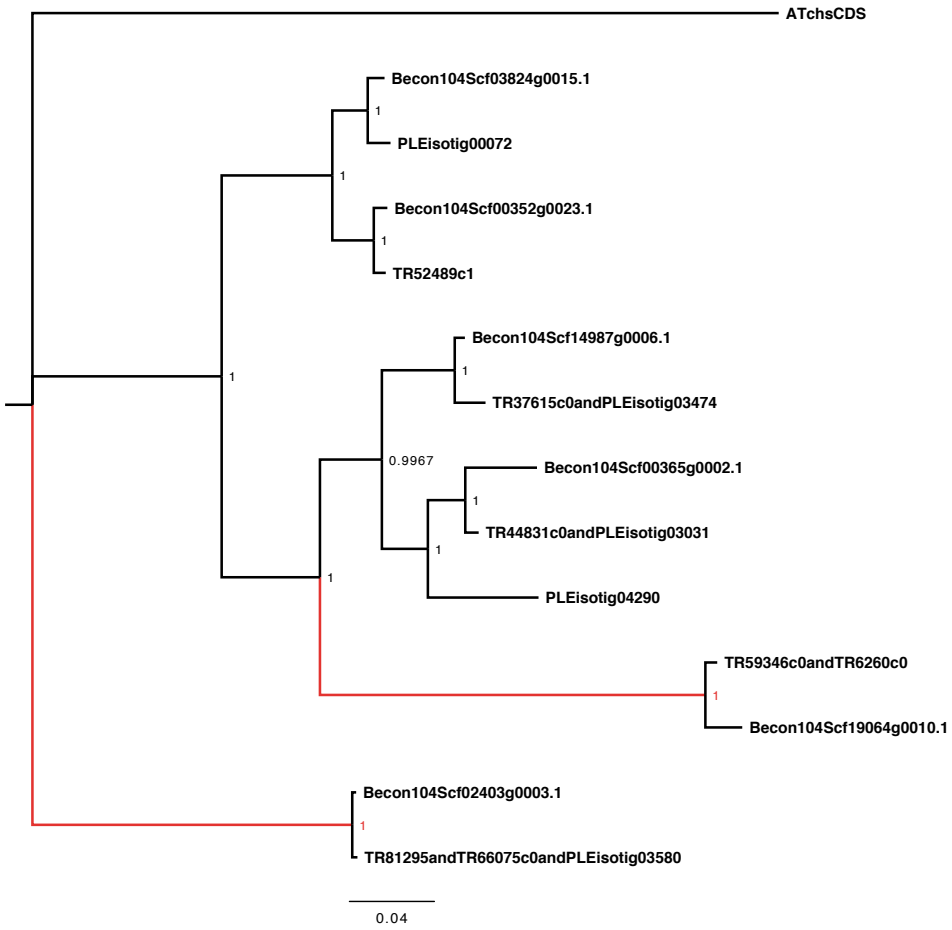



Figure 3. Gene tree of CHS sequences identified in *B. conchifolia* and *B. plebeja*. Reconstructed using MrBayes using a 1,236 bp matrix, using *A. thaliana* CHS as an outgroup. Branches are in red if positive selection detected along them.

All branches in the CHS gene tree were tested for evidence of positive selection, only two branches showed significantly different dN/dS from the global dN/dS for the whole tree.

The branch leading to orthologs Becon104Scf19064g0010.1 from B. conchifolia and TR59346c0andTR6260c0 from *B. plebeja*, appeared to have a significantly different dN/dS from the remainder of the tree, suggesting the branch may be subject to positive selection. The long branch length evident from the gene tree (Figure 2, Figure 3) supports a scenario of a burst of rapid divergence after the duplication event, though the proximity of the two orthologs suggests that this happened before the speciation of *B. conchifolia* and *B. plebeja*.

The second branch shown to have significant deviation in dN/dS relative to the global dN/dS of the tree leads to the orthologs Becon104Scf02403g0003.1 in *B. conchifolia* and TR81295andTR6607candPLEisotig03580. The ortholog pair are derived from the first duplication in the gene family following divergence from *A. thaliana*, and lack any duplications since their origin.

4.3.6 Expression of *Begonia* chalcone synthase

| | 00365g0002.1 | | 03824g0015.1 | | 00352g0023.1 | | 02403g0003.1 | | 14987g0006.1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C | P | C | P | C | P | C | P | C | P |
| ♀ | 10.5 | 12.3 | 202.2 | 112.8 | 1168.2 | 1450.2 | 180.1 | 190.4 | 47.7 | 55.9 |
| L | 17.0 | 3.2 | 329.5 | 174.0 | 1749.5 | 1936.5 | 310.1 | 54.3 | 69.1 | 17.8 |
| ♂ | 2.9 | 41.7 | 30.8 | 41.1 | 2220.2 | 5133.2 | 25.8 | 12.1 | 8.9 | 225.4 |
| P | 11.4 | 2.3 | 142.8 | 50.2 | 574.1 | 651.0 | 18.7 | 13.2 | 27.0 | 15.0 |
| R | 73.0 | 167.2 | 570.4 | 573.9 | 2326.7 | 2546.7 | 410.7 | 487.0 | 220.2 | 364.1 |
| V | 33.2 | 39.6 | 629.9 | 477.1 | 2321.3 | 2377.0 | 486.9 | 554.8 | 170.6 | 153.6 |

Table 4. Relative expression of *B. conchifolia* and *B. plebeja* CHS loci (denoted C or P) in all six tissues tested (♀= female flower, L = leaf, ♂ = male flower, P = petiole, R = root, V = vegetative bud.

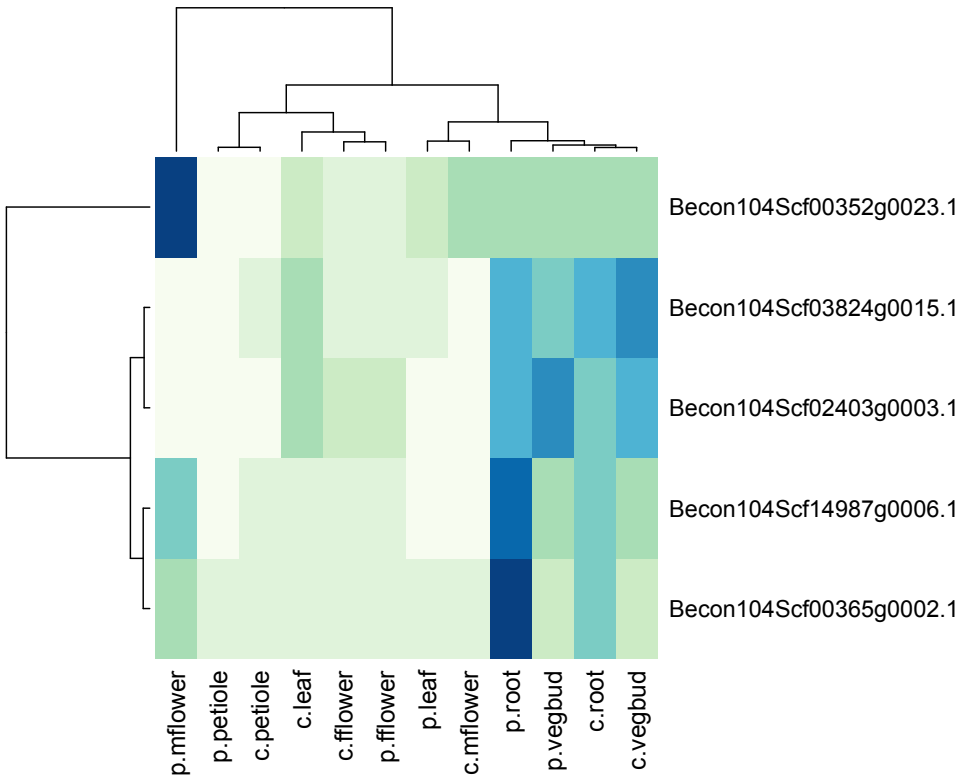



Figure 4. Heatmap of expression similarity across tissues and CHS loci. Distances are calculated using Euclidian distance and clustered using hierarchical clustering. Darker colours indicate higher relative expression, light colours indicate lower relative expression.

Expression data was obtained for 5 of the 6 identified members of the CHS gene family in *Begonia*. Becon104Scf19064g0010.1, omitted from expression analyses, had very low count data, rendering it unusable in EdgeR normalization. In both species, one copy of CHS (Becon104Scf00352g0023.1) appears to be dominant in all species, having an up to 21 fold increase in expression from the next highest expressed copy in the same tissue (Table 4, Figure 5). The pattern of expression of Becon104Scf00352g0023.1 is remarkably similar between *B. conchifolia* and *B. plebeja*, with the exception of male flower, where *B. plebeja* has an almost 2 fold increase in expression. In contrast to the male flower, the expression profiles of female flower in *B. conchifolia* and *B. plebeja* were very similar (Figure 4, Figure 5).

Interestingly, root and vegetative buds generally tended to show the highest expression in all CHS copies excluding Becon104Scf00352g0023.1. Vegetative buds showed conserved expression across

3 of the 5 CHS copies, with some evidence of divergent expression in Becon104Scf03824g0015.1 and Becon104Scf02403g0003.1 (Table 4). *B. plebeja* roots showed upregulation of Becon104Scf02403g0003.1, Becon104Scf14987g0006.1 and Becon104Scf00365g0002.1 relative to *B. conchifolia*, and both species showed upregulation of Becon104Scf00365g0002.1 in root relative to all the other tissues. The removal of the dominant Becon104Scf00352g0023.1 CHS copy allowed observation of patterns of the copies with lower expression (Figure 5). An immediately noticeable difference was the near absence of Becon104Scf 14987g0006.1 in *B. conchifolia* male flower, and the much higher expression in the same tissue in *B. plebeja*. Becon104Scf03824g0015.1 and Becon104Scf02403g0003.1 showed an interesting pattern of divergence between *B. conchifolia* and *B. plebeja*, most prominently in tissue pairs female flower and leaf, and root and vegetative bud. *B.conchifolia* shows similar patterns of expression between the two CHS copies in both tissue pairs respectively, a trend summarised by the higher expression of Becon104Scf03824g0015.1 than Becon104Scf02403g0003.1, more so in root and vegetative bud than female flower and leaf. In these tissues, *B. plebeja* shows less congruent patterns, with reciprocal up- and down-regulation acting on each tissue pair for the two CHS copies. Where Becon104Scf03824g0015.1 was the dominant copy in *B. conchifolia* in the four tissues discussed here, in *B. plebeja* female flower and leaf showed Becon104Scf02403g0003.1 and Becon104Scf03824g0015.1 to be dominant in each tissue respectively, and the same trend was observed in root and vegetative bud.

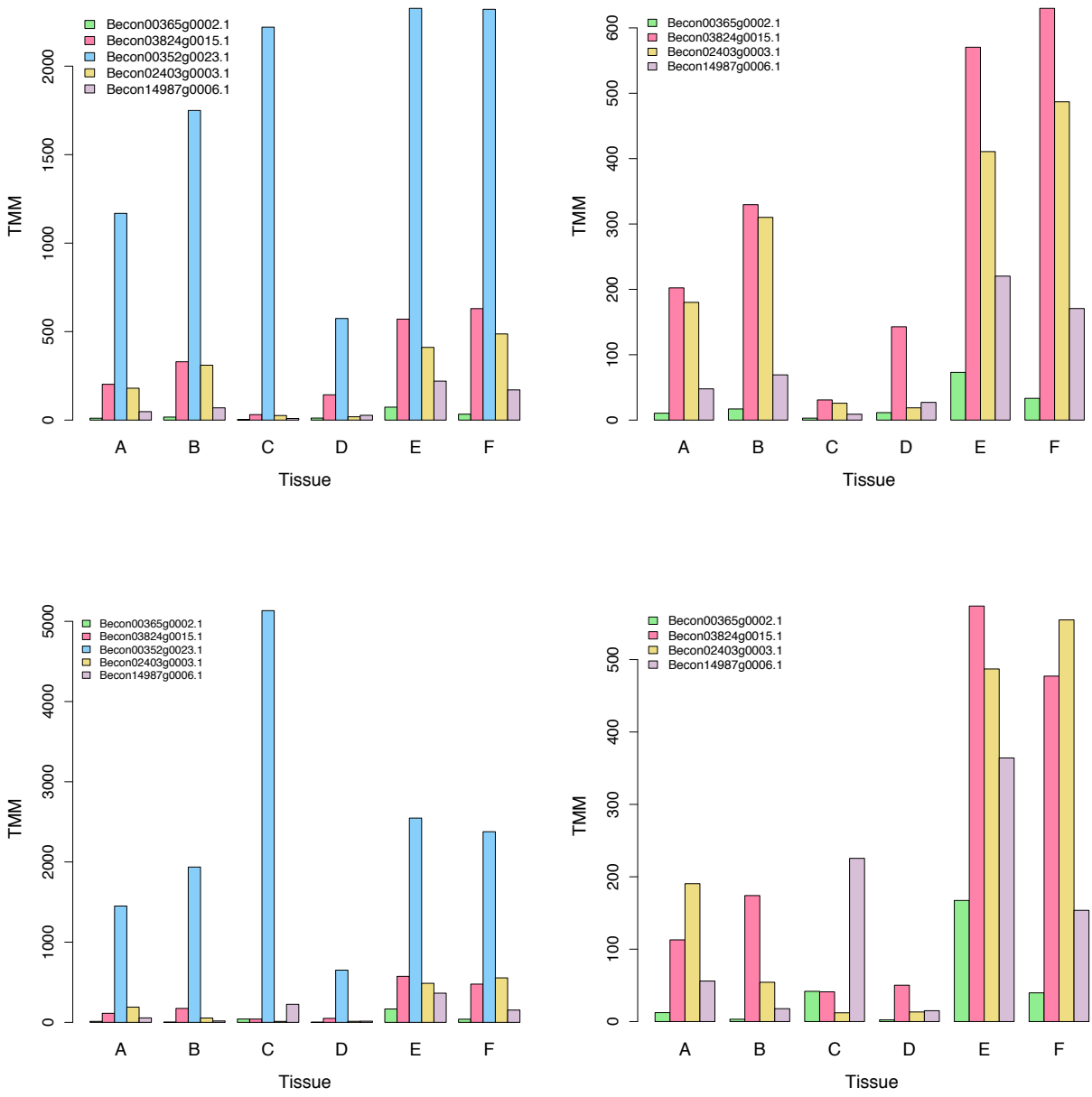



Figure 5. Top left: relative expression of CHS copies in *B. conchifolia*. Top right: relative expression of CHS copies excluding Becon00352g0023.1 to visualise details of less expressed copies. Bottom row: same as for top row but in *B. plebeja*.

## 4.4 Discussion

### 4.4.1 High variation in CHS copy number found in *Begonia*

A PCR survey of a range of South American species and one South East Asian species from *Begonia* revealed a range of CHS copy numbers within one genus. Species surveyed from South America have at least 6 copies of CHS, while Asian *B. puspitae* has a minimum of 5. *B. odorata* has the most duplicates, with 11 copies of CHS in total which, compared to the single CHS copy in the closest sequenced relative *C. sativus,* points to a recent burst of CHS duplications in the the genus *Begonia.* *A. thaliana*, like *C. sativus,* has only one copy of CHS, and this has been found to be the case for a large number of species closely related to *A. thaliana* (Koch et al. 2000), suggesting it is likely this group of related species has a lower rate of duplication, or CHS duplicates are selectively disadvantageous in that lineage, making the case of *Begonia* and the widespread retention of CHS duplicates in the genus very interesting, especially from an ecological perspective.

Labile gene families have been found in the genus *Gossypium,* (Small and Wendel, 2000) where the Alcohol Dehydrogenase (Adh) gene family, while having been shown have a conserved copy number in many closely related species, had variable copy number when examining the *Gossypium* genus at high resolution. A complex pattern of duplication observed in *Gossypium*, with different duplicates originating at markedly different time points in evolution, stands in contrast to the burst of duplication seen in *Begonia*; all duplicates sequenced using Sanger sequencing as well as the RNA-seq derived CHS sequences from *B. conchifolia* and *B. plebeja* form a monophyletic group originating after the divergence of *Begonia* from *C. sativus*.

Rapid duplication within gene families far exceeding average duplication and fixation rates for mammals (1 every 7 million years) (Lynch and Conery, 2001) have been found in both mammals (Laukaitis et al. 2008) and plants (Kong et al. 2007), with tandem duplication and non-allelic homologous recombination playing a large role in the duplication history of the gene families. While synteny data is unavailable to identify the mechanism of the *Begonia* duplications, it is likely that they arose by means other than WGD (tandem, non-allelic homologous recombination, retrotransposition) based on the pattern of recent and rapid duplication, which is not mirrored by corresponding chromosome numbers in the same species.

### 4.4.2 Evidence of positive selection in *Begonia* chalcone synthase

Maximum likelihood analysis of CHS coding sequence in *B. conchifolia* and *B. plebeja* duplicates revealed two branches, both leading to an orthologous pair of duplicates in both species, to have signatures of positive selection. The ages of the duplications were strikingly different; the first duplication after the divergence of *Begonia* from *A. thaliana*, which was used as an outgroup, showed signs of positive selection, though the time during which nonsynonymous mutations began to accumulate on this branch is not clear. Therefore, though the branch is basal in the CHS gene tree, selection may not have acted on it immediately after duplication, but more recently. The second branch under positive selection is derived from a more recent duplication, and has a longer branch length than adjacent branches, indicating that both synonymous and nonsynonymous mutations are being accumulated on this branch. No data is available currently to determine whether selection has caused any functional changes such as shifts in substrate specificity, *in vitro* analyses of catalytic rates in the presence of different substrates will better determine to what extent the *B. conchifolia*

and *B. plebeja* duplicates have diverged in function, as well as whether this divergence occurred before or after the split of *B. conchifolia* and *B. plebeja*.

Previous investigations into the CHS gene family have identified varied patterns of gene duplication in CHS, associated with functional shifts in substrate and product, prompting the suggestion of defining a CHS superfamily (Schröder, 1997). This observation supports a hypothesis that duplication in the plant CHS gene family promotes adaptive evolution of new duplicates to catalyse new substrate forms. Studies in *Ipomoea* have shown the CHS gene family in this genus to comprise 5 copies, phylogenetically forming two distinct groups termed ABC and DE (Yang et al. 2004). The authors showed that the group termed ABC, as well as having a higher nonsynonymous mutation rate than DE, had diverged in substrate specificity; DE still using naringenin chalcone, while ABC used bisnoryangonin. Interestingly, much of the coding sequence of the CHS duplicates was found to be under purifying selection, suggesting that relatively few mutations, such as at catalytically important sites such as the coumaroyl binding pocket are enough to change substrate specificity.

The finding that duplication in CHS promotes adaptive evolution to create greater functional diversity is supported elsewhere; comparison of CHS duplicates in the common ancestor of *Dendranthema* and *Gerbera* showed that positive selection acting on one subfamily of CHS also caused functional divergence, creating new CHS like genes in both species which have altered substrate preference (Yang et al. 2002).

Different duplication ages have been shown to accumulate mutations differently; Han et al. (2014) showed that duplications in CHS and CHS like genes across 13 plant families formed distinct clades of early diverged and late diverged duplicate CHS. While synonymous mutation rates were lower in the late diverged and more recent duplicates, this group also showed higher relative rates, suggesting that these later CHS and CHS like sequences were more predisposed to adopt new functions. It is

interesting therefore that both younger and older duplications in the *Begonia* CHS gene tree have branches showing positive selection. The relatively recent burst of duplication in *Begonia* CHS may explain this pattern; though the two branches found to be under positive selection are of different ages, both are recent relative to the age of the genus. It is worth noting that plant anthocyanin diversity has benefitted from gene duplication in gene families other than CHS; MYB transcription factor duplicates were found to be under positive selection in *Scutellaria*, functional annotation of the duplicates revealed these were transcription factors regulating genes of the anthocyanin biosynthesis pathway (Huang et al. 2015). The authors speculated that these transcription factors may underlie flower colour changes seen between *Scutellaria* species associated with pollinator shifts.

### 4.4.3 Divergent trends of expression in lower expressed copies of chalcone synthase

Comparison of relative expression of different copies of CHS across six tissues in *B. conchifolia* and *B. plebeja* revealed one copy of CHS in both species to be the foremost source of CHS transcripts in all tissues sampled. Interestingly, the pattern of expression across tissues appears to be conserved between *B. conchifolia* and *B. plebeja*, except for a markedly higher expression of the dominant copy in *B. plebeja* male flower, having around a two-fold increase in expression compared to it's counterpart in *B. conchifolia*. Divergent patterns of expression are apparent in the copy pair Becon104Scf03824g0015.1 and Becon104Scf02403g0003.1; the pattern of reciprocal dominance of the two copies in tissue pairs female flower and leaf, and root and vegetative bud in *B. plebeja* is absent in *B. conchifolia*, where the pattern of expression within the aforementioned tissue pairs was similar.

### 4.4.4 Conclusion

This chapter has examined more in depth the patterns evident in the Chalcone Synthase gene family, one that was identified as a candidate for expressional divergence in Chapter 3. The duplication and expression patterns in this gene family suggest that Chalcone Synthase is a highly labile family in the genus *Begonia*. Though it is not yet clear whether the variable copy number seen in the genus is a result of preferential retention or rapid fixation via high genetic drift conferred by restricted populations, the evolutionary consequences of such differential duplication could be important in *Begonia* phenotypic diversity. Analysis in this chapter indicate that already changes in expression and selection pressure have created divergent patterns between *B. conchifolia* and *B. plebeja*, though further work must be done to understand the effect of these changes on the phenotype and interaction with the environment of the study species.

# Chapter 5: Conclusions

This thesis sets out to determine whether gene duplication may have played a role in *Begonia* speciation by answering three questions:

- Are gene duplications common?
- Do the fates of duplicated genes suggest neofunctionalisation?
- Do the patterns of gene duplication and changes in expression indicate drivers of speciation?

While investigating patterns of diversity on a genome wide scale across the taxonomic range of *Begonia*'s near 2000 species is currently unfeasible, a highly suitable model was found in two South American species from section Gireoudia, *B. conchifolia* and *B. plebeja.* The two species, while being very closely related and interfertile, are phenotypically highly divergent and occupy different habitats. Such patterns of phylogenetic closeness and phenotypic divergence, as well as evidence from previous studies of a whole genome duplication being shared between these two species (Brennan et al. 2012), provides a good system to ask how species in *Begonia* have come to be so phenotypically divergent, and how, if at all, gene and genome duplication may have helped facilitate this. Results of the thesis have suggested that gene and genome duplication has had a significant impact on the divergence of *B. conchifolia* and *B. plebeja*, and may have effects on *Begonia* phenotypic and ecological diversification on a wider scale.

## 5.1 Are gene duplications common?

*B. conchifolia* showed comparatively low levels of duplication in comparison with *A. thaliana* and *G. max* (Chapter 3), both of which have had undergone whole genome duplications at least once in their evolutionary history (Clarindo et al. 2007, Simillion et al. 2002), appearing to have similar patterns of duplication to the much closer related *C. sativus*, which lacks a recent whole genome duplication (Huang et al. 2009).

The finding that *B. conchifolia* had a relatively unduplicated genome sheds some light on the evolutionary processes that may have been at play after the divergence of lineages leading to *A. thaliana* and *Begonia*. A genome duplication identified at the base of the *Begonia* lineage between 24 to 45 MYA, or in the most recent common ancestor of *Begonia* and *Hillebrandia sandwichensis*, the sister genus to *Begonia*, (Brennan et al. 2012), provides evidence that *B. conchifolia* does indeed have a relatively recent large scale duplication.

The lack of a recent genome duplication in *C. sativus* supports the evidence that the identified genome duplication happened near the emergence of the *Begonia* genus.

Recent whole genome duplication is not incongruent with patterns seen across sections of *Begonia*; the high variability of C-value and chromosome number between sections is indicative of rapid genome evolution, with large scale duplications and chromosome fusion and fission events being relatively frequent (DeWitte et al. 2009).

While this is the case, the genomic variation is different across sections, some having much greater differences than others, and section Gireoudea, which holds both *B. conchifolia* and *B. plebeja*, has markedly lower variation in both chromosome number and size than sections such as Platycentrum, Gaerdtia and Begonia. For example, the C-values within section

Gireoudea range between roughly 0.7 and 0.8, whereas a section such a Gaerdtia ranges from around 0.5 to 1.1 (Figure 1, Chapter 3). Genome downsizing has likely played a role in the loss of gene duplicates derived from the genome duplication at the base of *Begonia*. Given that either all or the majority of *Begonia* species share the genome duplication, it is likely that genome downsizing occurred relatively quickly, though it is not clear whether most of it occurred before species diversification began in the ancestral *Begonia*.

If not, most species will have inherited a greatly reduced genome, with the loss of the majority of whole genome duplication derived sequence, and the variation seen in C-value and chromosome number is the result of lineage specific genomic events.

## 5.2 Do the fates of duplicated genes suggest neofunctionalisation?

The neofunctionalisation of duplicate genes is most readily achieved by the divergence of expression profiles, allowing duplication in such cases to be a highly potent force for facilitating new phenotypes through regulatory change (Harikrishnan et al. 2015, Gao et al. 2015). Studies in *Brassica rapa* have shown divergent expression patterns in highly duplicated Auxin Response Factor and Auxin/Indole Acetic Acid gene families (Huang et al. 2015). The authors showed that the two gene families were differentially expanded, and higher levels of duplicate gene retention were associated with more tissue specific expression patterns. Overall patterns of expression in *B. conchifolia* and *B. plebeja* (Chapter 3, Figure 4) agreed with the patterns identified in Huang et al. (2015) revealing that larger gene families did indeed have greater diversity of expression, indicating that this phenomenon is not restricted to just a handful of genes.

A study in *Populus euphratica* (Brinker et al. 2010) lends support to the importance of context in the expressional evolution of duplicate genes; investigating the Heat Shock Factor gene family in *P. euphratica* revealed specialization of duplicates to different stress responses. Comparison of the divergence of the same gene family in other plant species such as *A. thaliana* revealed different patterns of specialization consistent with species specific stressors. Divergence in duplicate gene expression in one *Begonia* species but not another supports such studies, indicating that, while duplication of genes may promote new functions being assumed by duplicates, context specific responses in duplicate gene neofunctionalization result when divergent selective pressures are exerted on a duplicate pair in two species or populations. Two examples explored further in Chapter 3 showed divergent expression between duplicates in either *B. conchifolia* and not *B. plebeja*, or vice versa. Chalcone Synthase, a type III polyketide synthase in the anthocyanin biosynthetic pathway, showed modest expression conservation between a duplicate pair in *B. conchifolia*, with a Pearson correlation coefficient of 0.65, while the same gene pair in *B. plebeja* showed almost no similarity with a correlation coefficient of -0.08. Conversely, 3-ketoacyl CoA synthase, a gene involved with wax cuticle synthesis, had two gene duplicate pairs which showed divergent conservation of expression, this time being highly conserved in *B. plebeja* (Pearson correlation coefficients of 0.97 and 0.96 respectively) and divergent in *B. conchifolia* (Pearson correlation coefficients of 0.07 and 0.15 respectively).

5.3 Do the patterns of gene duplication and changes in expression indicate drivers of speciation?

Functional annotation and enrichment analysis of gene duplicates which were differentially conserved in expression pattern between *B. conchifolia* and *B. plebeja* identified a modest number of GO categories which were enriched in the gene set of interest (Chapter 3, Table 2). Identification of the genes underlying functional categories significantly enriched revealed that a small number of genes were responsible, of most interest based on adaptive value being the key anthocyanin biosynthetic pathway gene Chalcone Synthase, and 3-Ketoacyl-CoA Synthase, a gene involved in fatty acid elongation, and more specifically, cuticular wax synthesis. These two genes may reveal some of the drivers of divergence between *B. conchifolia* and *B. plebeja*.

3-Ketoacyl-CoA Synthase is involved with the response of plants to drought, being shown to be upregulated highly under simulated drought conditions in *Cynanchum komarovii*, a xerophytic species distributed across China (Ma et al. 2015), and dynamic build up of cuticle in response to water deficit was seen in *A. thaliana* plants (Kosma et al. 2009).

One of the main differences in habitat between *B. conchifolia* and *B. plebeja* is the exposure of heat and moisture; the former living in shaded habitats where moisture is more abundant as a result of more shading, and the latter living in a more open and drought prone environment, making moisture less readily retainable. This key environmental difference suggests *B. plebeja* has a much higher investment in drought tolerance and prevention of water loss through non-stomatal surfaces.

The leaf cuticle, however, also has a role in defence against pathogenic attack; as well as forming a physical barrier to fungal attack, the plant senses cutinases released by pathogenic fungi, allowing the plant to launch a defence response. However, fungal spores, similarly, are able to recognise contact with cuticle tissue, allowing spores to begin germination at an appropriate time to maximise chances of establishment (Serrano et al. 2014). Studies in barley used mutants for 3-Ketoacyl-CoA Synthase 6 to show that mutants with lower cuticle deposition had less frequent fungal germination events, providing further support for this mechanism (Weidenbach et al. 2015). The moister, shaded and cool habitat of *B. conchifolia* compared to *B. plebeja* suggests that pathogen attack, especially fungal, may be more prevalent here, exposing *B. conchifolia* to selective pressures to avoid cuticle tissue recognition by fungal pathogens. The longer lived leaves of *B. conchifolia* will also repay investment in defense better than the leaves of *B. plebeja* which usually last only for one season. Chalcone Synthase, the other gene selected as a putative candidate alongside 3-Ketoacyl-CoA Synthase, may underlie another of the contrasting ecological features between the habitats of *B. conchifolia* and *B. plebeja*; exposure to light. As a key enzyme in the anthocyanin biosynthetic pathway, Chalcone Synthase is important for the production of many secondary compounds many of which have roles in protection from UV damage, but also have myriad other roles in defence, such as free radical scavenging and deterrence of herbivores (Manetas et al. 2002, Schaefer and Rolshausen, 2006). Adaptation to light levels in the long term is important for appropriate investment of resources at the correct time; annual patterns such as colder days and shorter light regimes act as cues for plants to increase levels of epidermally deposited anthocyanins, as well as reduce the amount carbon invested into lignin production, reducing the rate of plant growth. This phenomenon, shown by Zhang et al (2016) in *Begonia semperflorens* showed that the interplay of environmental cues is

responsible for this coordinated response, and the absence of one cue, such as short-day light regime, can prevent the timely response of *B. semperflorens* to autumnal conditions.

As well as seasonal cues, success of individuals across their life as measured by efficient growth and utilisation of resources is highly dependent on the fine-tuning of photosynthetic and photoprotective activity to suit the light availability. Krause et al. (2012) found that some shade tolerant species had highly reduced growth rates when grown in high light conditions, though this was not a pattern replicated across all species tested, suggesting that photoprotective responses in some plants are adapted to a given range of UV exposure expected in the native habitat. Finally, short term roles of photoprotection conferred by anthocyanins are particularly pertinent in tropical understorey species such as *B. conchifolia*. While *B. plebeja* has greater sun exposure over longer periods, tropical understorey species like *B. conchifolia* rely greatly on sunflecks produced by transient openings in the canopy roof (Chazdon, 1988, Chazdon and Pearcy, 1991). Utilising this short burst of UV radiation efficiently requires prevention of the photoinhibitive effects of high UV on photosystem II efficiency (Hughes et al. 2014). Though not constituting a uniform response among species tested, Barnes et al. (2016) found a number of species to display very rapid mobilisation of flavonoid compounds in response to high UV exposure, suggesting the evolution of mechanisms to deal with sudden bursts of UV radiation such as sunflecks. The large gene families of CHS identified in *Begonia* suggest both a high rate of gene duplication and high retention of copies. Similar lineage specific duplication patterns have been seen in other taxa; Zavala and Opazo (2015) identified lineage specific expansions in the Chalcone Synthase gene family in the Fabales, suggesting selection for greater investment in secondary compounds for defence.

In the case of *Begonia,* the defence may be against the abiotic stress of high light intensities rather than fungi and herbivores. A mapping population of *B. conchifolia* and *B. plebeja* used to construct a genetic map is available and can be used to test this, as anthocyanin patterns are highly variable and have been linked to two QTLs on chromosome 9. The mapping population also could provide a means to test a link between the presence of anthocyanin in the leaves and resilience to photoinhibition, fungal attack or herbivores.

Some evidence of incipient reciprocal silencing was seen in *B. plebeja* Chalcone Synthase expression patterns, though not in *B. conchifolia* (Chapter 3). The relatively recent burst of duplication specific to *Begonia* giving rise to the CHS duplicates, along with evidence of expressional divergence is indicative of selective forces effecting the evolution of the CHS gene family as these species diverge. While it is tempting to speculate that the expansion of the Chalcone Synthase gene family allowed the adaptation of the two study species to different light levels, there is need for further investigation of expressional responses of Chalcone Synthase duplicates in the two species in response to different levels of UV radiation. The mapping population also provides a way to examine the expression of different paralogs in different genetic backgrounds, and under different light levels.

## 5.4 Genomic analysis in non-model organisms

The decreasing cost of sequencing has provided a new tool to understand biological diversity through large scale sequence analysis. The generation of tissue-specific transcriptomes for two closely related species of *Begonia* has allowed us to use comparative sequence analysis and expression analysis to determine patterns of evolution. The RNA-seq data has also allowed robust annotation of the draft *Begonia conchifolia* genome (Bombarely et al.

Unpublished), allowing its use as a reference genome for this mega diverse genus, which is key for further development of evolutionary studies in this group. As well as a reference for further comparative genomics, it has also been used to design a set of hybrid baits for enrichment of selected sequences, allowing phylogenetic and molecular evolutionary analyses in *Begonia* (Cho et al. unpublished). The analyses presented here have highlighted some of the difficulties working with non-model organisms and limited material/depth of sequence, thus providing guidance for such future studies. Obtaining a good reference assembly is vital to robust analyses downstream, and the testing of different assembly strategies proved to be an invaluable tool to selecting the best assembly in terms of completeness and accuracy of transcript reconstruction. Optimisation of transcriptome assembly has shown that large differences are apparent between different strategies, and highlights the importance of factors such as number, length and quality of reads in the input FASTQ files, and therefore the complexity of transcript models predicted by assembly software such as Trinity (Grabherr et al. 2011).

De novo assembly software algorithms are diverse, and the different strategies used by them produces sometimes highly variable results. Chen et al. (2015) showed a large number of transcripts reconstructed were unique to a specific assembler, illustrating the extent to which different methods deliver benefits to different classes of gene structures.

Most of the frequently used *de novo* transcriptome assemblers, such as Trinity (Grabherr et al. 2011), TransAbyss (Robertson et al. 2010), and Shannon (Kannan et al. 2016) use a method utilising de Bruijn graphs. Where de Bruijn graphs are used in genome assembly, the representation of sequence is relatively uniform. In the case of *de novo* transcriptome assembly, differential expression leads to highly variable sequencing depths across transcripts, makes de Bruijn graph assembly more error prone (Robertson et al. 2010).

Some studies have made use of hybrid assemblies, using reads generated by short and long read technologies. Cahais et al. (2012) combined reads from Illumina and Roche 454 sequencing platforms to test whether longer reads improved transcript recovery and misassembly rates, finding that even small amounts of long read data such as Roche 454 reads improved assembly metrics. Using longer read sequencing for transcriptome, as well as genome assemblies has been treated with caution; both Roche 454 and Pacific Biosciences (PacBio) technologies introduce much higher error rates than shorter read technologies such as Illumina (Quince et al. 2009, Quail et al. 2012). Read correction strategies have been developed in order to address this, and results of validation studies have found markedly improved results in both PacBio and Illumina reads (Nikolenko et al. 2013, MacManes and Eisen, 2013). The increase in the use of long read technologies such as PacBio and Nanopore will likely help to resolve many of the issues involved with assembly issues, and improvements in error rates and development of error correction software makes this a highly suitable option for accurately reconstructing full length isoforms. Other developments in de novo transcriptome assembly methods have addressed problems in currently used methodologies such as de Bruijn graphs. The *de novo* transcriptome assembly software Shannon has aimed to address the effects of repetitive sequences introduced by alternative transcripts and low coverage in rare transcripts by dynamically optimising the kmer length during assembly. Kmer length is increased in areas of high repetitiveness in order to help resolve these regions, and shortening kmer length in low coverage regions maintains graph connectivity (Kannan et al. 2016). Alongside the development of new assembly methods, strategies such as the tr2aacds pipeline in the EvidentialGene software suite (Gilbert, 2013) used in Chapter 3 are helpful for circumventing issues associated with individual *de novo* assemblers by combining non-redundant transcripts from multiple assemblers.

A good assembly is only part-way to a useful data set for comparative genomics. Key to the whole approach is correct assignment of relationships of orthology and paralogy to genes within and between species, however this remains very problematic and no fully robust solution has been devised (Dalquen et al. 2013).

First of all, identification of homologous sequence can be made difficult by portions of coding region being highly divergent. This is frequent in transcription factors where only the DNA and protein interaction domains are conserved to any extent. Promiscuous domains, such as those in some defence response proteins, have a similar confounding effect, though unlike unconserved regions which may mask signatures of homology, such domains can create homologous signal where there is none (Enright et al. 2002). These difficulties are amplified in the case of sequence data that is fragmented or misassembled due to insufficiently deep sequencing coverage. Some approaches to this problem acknowledge the difficulty in correctly assigning a read to a specific gene family member and opt instead to assign expression values to whole gene families (Robert and Watson, 2015).

Secondly, defining the relationships within gene families at high resolution has proven to be very difficult, chiefly due to complex duplication patterns, a very common feature of plant genomes in particular (Blanc and Wolfe, 2004). Missing data in the form of unrecovered loci due to low coverage sequencing can be problematic in cases where a gene family has complex duplication patterns, as this results in the misassignment of relationships (Dalquen et al. 2013). For example, using the bidirectional best hit method of assigning 1:1 orthology in taxa which are duplication rich results in as many as 60% of orthologous relationships being missed (Dalquen et al. 2013).

Polyploidy, and especially allopolyploidy, a hybridization event which is accompanied by a whole genome duplication, is common especially in plants, and defining orthologous and

paralogous relationships in such taxa is still challenging due to disparate origins of homeologous sequences (Glover et al. 2016). OrthoFinder (Emms and Kelly, 2015), a relatively recent approach designed to account for biases towards longer sequences as well as differential distances between input taxa, produced mixed results; while members of all gene families examined were correctly assigned based on manual inspection of alignments and BLAST searches against *A. thaliana*, more divergent members were found to be excluded and reported as single copy genes. More divergent gene families may lie outside the scope of the threshold parameters used in OrthoFinder; the implications of this indicate that next steps in homology assignment must address the problematic 'one size fits all' approach.

## 5.5 Further work

This thesis has produced resources and devised strategies to uncover candidate genes which may have been targets of selection during the speciation of *B. conchifolia* and *B. plebeja*. More work is required, however, to shed more light on the exact phenotypic changes associated with expressional divergence of candidate genes in these two species. Expression data identified for candidate genes must be verified by qRT-PCR to confirm these expression patterns. After confirmation of expression in normal conditions, expression data can be obtained for the candidate genes in the six tissues sampled here, and additionally under a range of environmental conditions. The gene expression measurements of identified candidate gene families 3-Ketoacyl CoA Synthase and Chalcone Synthase in *B. conchifolia* and *B. plebeja* under environmental stresses such as drought and high UV will help determine how different duplicates in these two candidate gene families are mobilised in response to stresses which are found in each species' habitat. The mapping population of *B. conchifolia*

and *B. plebeja* can be used to link specific duplicate loci to phenotypic traits, as well as conduct expression analyses to investigate promoter sequence variation.

Finally, the sequencing of the draft genome of *Hillebrandia sandwichensis*, the sister genus of *Begonia*, will reveal more about the duplication patterns of this closely related monotypic species, as well as allow a better estimation of when the whole genome duplication in *Begonia* occurred. Large differences in duplication patterns between species in *Begonia* and *H. sandwichensis* will give further support to the hypothesis that gene and genome duplication may have contributed to diversification of *Begonia*.

## 5.6 Conclusions

Studying traits of importance and their contribution to the generation of diversity based on preconceived hypotheses can be an efficient way to uncover factors which have shaped the evolutionary history of taxa. Such strategies, however, can have limitations due to non-neutral expectations of biotic and abiotic forces important in shaping plant diversity. Identifying phenotypes based on underlying genetic changes driven by selection, however, can uncover traits important in the divergence of species that are not immediately apparent, and may be missed by hypothesis driven studies.

With such candidate gene finding studies, while false negative and false positive results are an inherent risk, careful validation of phenotype-genotype relationships of preliminary candidates allows a much more powerful examination of the whole of the plant's biology, as well as the genetic basis of diversity.

# Bibliography

Adams. K. L, Wendel. J. F, 2005, Polyploidy and genome evolution in plants, Current opinion in plant biology, **8**(2), 135-41

Altschul. S. F, Gish. W, Miller. W, Myers. E. W, Lipman. D. J, 1990, Basic local alignment search tool, Journal of molecular biology, **215**(3), 403-410.

Alves-Carvalho. S, Aubert. G, Carrère. S, Cruaud. C, Brochot. A, Jacquin. F, Klein. A, Martin. C, Boucherot. K, Kreplak. J, da Silva. C, Moreau. S, Gamas. P, Wincker. P, Gouzy. J, Burstin. J, 2015, Full-length de novo assembly of RNA-seq data in pea (Pisum sativum L.) provides a gene expression atlas and gives insights into root nodulation in this species, The plant journal, **84**(1), 1-19

Ambavaram. M. M. R, Basu. S, Krishnan. A, Ramegowda. V, Batlang. U, Rahman. L, Baisakh. N, Pereira. A, 2014, Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress, Nature communications, **5**, 5302

Archetti. M, Brown. S. P, 2004, The coevolution theory of autumn colours, Proceedings of the Royal Society B, **271**(1545), 1219-23

Archetti. M, 2009, Evidence from the domestication of apple for the maintenance of autumn colours by coevolution, Proceedings of the Royal Society B, **276**(1667), 2575-80

Aury. J, Jaillon. O, Duret. L, Noel. B, Jubin. C, Porcel. B. M, Ségurens. B, Daubin. V, Anthouard. V, Aiach, N, Arnaiz. O, Billaut. A, Beisson. J, Blanc. I, Bouhouche. K, Câmara. F, Duharcourt. SGuigo. R, Gogendeau. D, Katinka. M, Keller. A, Kissmehl. R, Klotz. C, Koll. F, Le Mouël. A, Lepère. G, Malinsky. S, Nowacki. M, Nowak. J. K, Plattner. H, Poulain. J, Ruiz. F, Serrano. V, Zagulski. M, Dessen. P, Bétermier. M, Weissenbach. J, Scarpelli. C, Schächter. V, Sperling. L, Meyer. E, Cohen. J, Wincker. P, 2006, Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia, Nature, **444**(7116), 171-8

Barakat. A, DiLoreto. D. S, Zhang. Y, Smith. C, Baier. K, Powell. W. A, Wheeler. N, Sederoff. R, Carlson. J. E, 2009, Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection, BMC plant biology, **9**, 51

Barnes. P. W, Tobler. M. A, Keefover-Ring. K, Flint. S. D, Barkley. A. E, Ryel. R. J, Lindroth. R. L, 2016, Rapid modulation of ultraviolet shielding in plants is influenced by solar ultraviolet radiation and linked to alterations in flavonoids, Plant, cell and environment, **39**(1), 222-30

Bennett. M, 2004, Biological relevance of polyploidy: ecology to genomics Perspectives on polyploidy in plants – ancient and neo, Biological journal of the Linnaean Society, **82**, 411-23

Berardini. T. Z, Reiser. L, Li. D, Mezheritsky. Y, Muller. R, Strait. E, Huala. E, 2015, The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome, Genesis, **53**(8), 474-85

Berens. A. J, Hunt. J. H, Toth. A. L, 2015, Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects, MBE, **32**(3), 690-703

Bi. K, Vanderpool. D, Singhal. S, Linderoth. T, Moritz. C, Good. J. M, 2012, Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales., BMC genomics, **13**, 403

Biscotti. M. A, Gerdol. M, Canapa. A, Forconi. M, Olmo. E, Pallavicini. A, Barucca. M, Schartl. M, 2016, The Lungfish Transcriptome: A Glimpse into Molecular Evolution Events at the Transition from Water to Land, Scientific reports, **6**, 21571

Blacklock. B. J, Jaworski. J. G, 2006, Substrate specificity of Arabidopsis 3-ketoacyl-CoA synthases, Biochemical and biophysical research communications, **346**(2), 583-90

Blanc. G, Wolfe. K, 2004, Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes, **16**, 1667-78

Boetzer. M, Henkel. C. V, Jansen. H. J, Butler. D, Pirovano. W, 2011, Scaffolding pre-assembled contigs using SSPACE, Bioinformatics, **27**(4), 578-79

Boetzer. M, Pirovano. W, 2014, SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information, BMC bioinformatics, **15**(1), 211

Bolger. A. M, Lohse. M, Usadel. B, 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics, **30**(15), 2114-20

Boucher. F. C, Casazza. G, Szövényi. P, Conti. E, 2016, Sequence capture using RAD probes clarifies phylogenetic relationships and species boundaries in Primula sect. Auricula, MPE, **104**, 60-72

Bowers. J. E, Chapman. B. A, Rong. J, 2003, Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events, Letters to nature, **422**, 433-38

Brennan. A. C, Bridgett. S, Shaukat Ali. M, Harrison. N, Matthews. A, Pellicer. JTwyford. A. D, Kidner, C. A, 2012, Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, Begonia, Tropical plant biology, **5**(4), 261-76

Brereton. N. J. B, Gonzalez. E, Marleau. J, Nissim. W. G, Labrecque. M, Joly. S, Pitre. F. E, 2016, Comparative Transcriptomic Approaches Exploring Contamination Stress Tolerance in Salix sp. Reveal the Importance for a Metaorganismal de Novo Assembly Approach for Nonmodel Plants, Plant physiology, **171**(1), 3-24

Brinker. M, Brosché. M, Vinocur. B, Abo-Ogiala. A, Fayyaz. P, Janz. D, Ottow.E. A, Cullmann. A. D, Saborowski. J, Kangasjärvi. J, Altman. A, Polle. A, 2010, Linking the salt transcriptome with physiological responses of a salt-resistant Populus species as a strategy to identify genes important for stress acclimation, Plant physiology, **154**(4), 1697-709

Brown. A. P, Kroon. J. T. M, Swarbreck. D, Febrer. M, Larson. T. R, Graham. I. A, Caccamo. M, Slabas. A. R, 2012, Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways, PloS one, **7** (2), e30100

Buggs. R. J. A, Chamala. S, Wu. W, Tate. J. A, Schnable. P. S, Soltis. D. E, Soltis. P. S, Barbazuk. B. W, 2012, Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin, Current biology, **22**(3), 248-52

Cahais. V, Gayral. P, Tsagkogeorga. G, Melo-Ferreira. J, Ballenghien. M, Weinert. L, Chiari. Y, Belkhir. K, Ranwez. V, Galtier. N, 2012, Reference-free transcriptome assembly in non-model animals from next-generation sequencing data, Molecular ecology resources, **12**(5), 834-45

Cao. H, Nuruzzaman. M, Xiu. H, Huang. J, Wu. K, Chen. X, Li. J, Wang. L, Jeong. J, Park. S, Yang. F, Luo. J, Luo. Z, 2015, Transcriptome Analysis of Methyl Jasmonate-Elicited Panax ginseng Adventitious Roots to Discover Putative Ginsenoside Biosynthesis and Transport Genes, International journal of molecular sciences, **16**(2), 3035-57

Cantarel. B. L, Korf. I, Robb. S. M. C, Parra. G, Ross. E, Moore. B, Holt. C, Alvarado. A. S, Yandell. M,
2008, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Research, **18**(1), 188-96

Chang. Z, Wang. Z, Li. G, 2014, The impacts of read length and transcriptome complexity for de novo assembly: a simulation study, PloS one, **9**(4), e94825

Chazdon.R. L, 1988, Sunflecks and their importance to forest understorey plants, Advances in ecological research, **18**, 1-63

Chazdon. R. L, Pearcy. R. W, 1991, The importance of sunflecks for forest understory plants, BioSciece, **41**(11), 760-66

Chen. Y, Zhang. Y, Yuan. S, Liu. H, Zeng. X, Zhang. H, 2015, Ethyl methane sulfonate induces disease resistance in Begonia × hiemalis Fotsch, Horticulture, environment and biotechnology, **55**(6), 498-505

Chen. S, McElroy. J. S, Dane. F, Peatman. E, 2015, Optimizing Transcriptome Assemblies for Leaf and Seedling by Combining Multiple Assemblies from Three De Novo Assemblers, The plant genome, **8**(1)

Chinwalla. A.T, Cook. L. L, Delehaunty. K. D, Fewell. G. A, Fulton. L. A, Fulton. R. S, Graves. T. A, Hillier. L. W, Mardis. E. R, McPherson. J. D, Miner. T. L, 2002, Initial sequencing and comparative analysis of the mouse genome. Nature, **420**(6915), 520-562.

Christensen. A. B, Gregersen. P. L, Schröder. J, Collinge. D. B, 1998, A chalcone synthase with an unusual substrate preference is expressed in barley leaves in response to UV light and pathogen attack, **37**, 849-57

Christoffels. A, Koh. E. G. L, Chia. J, Brenner. S, Aparicio. S, Venkatesh. B, 2004, Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes, MBE, **21**(6), 1146-51

Clarindo. W. R, de Carvalho. C. R, Alves. B. M. G, 2007, Mitotic evidence for the tetraploid nature of Glycine max provided by high quality karyograms, Plant systematics and evolution, **265**(1), 101-7

Conesa. C Madrigal. P, Tarazona. S, Gomez-Cabrero. D, Cervera. A, McPherson. A, Szcześniak. M. W, Gaffney. D. J, Elo. L. L, Zhang. X, Mortazavi. A, 2016, A survey of best practices for RNA-seq data analysis, **17**(1), 13

Cui. X, Lv. Y, Chen. M, Nikoloski. Z, Twell. D, Zhang. D, 2015, Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome, Molecular plant, **8**(6), 935-45

Dalquen. D. A, Dessimoz. C, 2013, Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals, GBE, **5**(10), 1800-6

Dao. T. T. H, Linthorst. H. J. M, Verpoorte. R, 2011, Chalcone synthase and its functions in plant resistance, Phytochemistry reviews, **10**(3), 397-412

De Marais. D, Rausher. M, 2008, Escape from adaptive conflict after duplication in an anthocyanin pathway gene, Nature, **454**(7205), 762-5

De Wilde. J. F. E, Hughes. M, Rodda. M, Thomas. D. C, 2011, Pliocene intercontinental dispersal from Africa to Southeast Asia highlighted by the new species Begonia afromigrata (Begoniaceae), Taxon, **60**(6), 1685-92

Degenkolbe. T, Do. P. T, Zuther. E, Repsilber. D, Walther. D, Hincha. D. K, Köhl. K. I, 2009, Expression profiling of rice cultivars differing in their tolerance to long-term drought stress, Plant molecular biology, **69**(1), 133-53

Dewitte. A, Leus. L, Eeckhaut. T, Vanstechelman. I, Van Huylenbroeck. J, Van Bockstaele. E, 2009, Genome size variation in Begonia, Genome, **52**(10), 829-38

Dewitte. A, Eeckhaut. T, Van Huylenbroeck. J, Van Bockstaele. E, 2010, Meiotic aberrations during 2n pollen formation in Begonia, Heredity, **104**(2), 215-23

Dewitte. A, Twyford. A. D, Thomac. D. C, Kidner. C. A, Van Huylenbroeck. J, 2011, The Origin of Diversity in Begonia : Genome Dynamism , Population Processes and Phylogenetic Patterns, The dynamical processes of biodiversity-case studies of evolution and spatial distribution, 27-52

Demuth. J, Hahn. M. W, 2009, The life and death of gene families., BioEssays, **31**(1), 29-39

Desjardins. C. A, Cerqueira. G, C, Goldberg. J. M, Dunning Hotopp. J. C, Haas. B. J, Zucker. J, Ribeiro. J. M. C, Saif. S, Levin. J. Z, Fan. L, Zeng. Q, Russ. C, Wortman. J. R, Fink. D. L, Birren. B. W, Nutman. T. B, 2013, Genomics of Loa loa, a Wolbachia-free filarial parasite of humans, Nature genetics, **45**(5), 495-500

Dlugosch. K. M, Lai. Z, Bonin. A, Hierro. J, Rieseberg. L. H, 2013, Allele identification for transcriptome-based population genomics in the invasive plant Centaurea solstitialis, G3, **3**(2), 359-67

Drummond. D. A, Bloom. J. D, Adami. C, Wilke. C. O, Arnold. F. H, 2005, Why highly expressed proteins evolve slowly, PNAS, **102**(40), 14338-43

Du. Z, Zhou. X, Ling. Y, Zhang. Z, Su. Z, 2010, agriGO: a GO analysis toolkit for the agricultural community, Nucleic acids research, **38**, 64-70

Durbin. M. L, Learn. G. H, Huttley. G. A, Clegg. M. T, 1995, Evolution of the Chalcone synthase family in the genus Ipomoea, PNAS, **92**, 3338-42

Durbin. M. L, McCaig. B, Clegg. M. T, 2000, Molecular evolution of the chalcone synthase multigene family in the morning glory genome, Plant molecular biology, **42**(1), 79-92

Eierman. L, Hare. M. P, 2015, Reef-Specific Patterns of Gene Expression Plasticity in Eastern Oysters (Crassostrea virginica), Heredity, esv057, 1-11

Emms. D, Kelly. S, 2015, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, Genome biology, **16**(1), 157

English. A. C, Richards. S, Han. Y, Wang. M, Vee. V, Qu. J, Qin. X, Muzny. D. M, Reid. J. G, Worley. K. C, Gibbs. R. A, 2012, Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology, **7**(11), 1-12

Enright. A. J, Van Dongen. S, Ouzounis. C. A, 2002, An efficient algorithm for large-scale detection of protein families, Nucleic acids research, **30**(7), 1575-84

Esteban. R, Fernández-Marín. B, Becerril. J. M, García-Plazaola. J. I, 2008, Photoprotective implications of leaf variegation in E. dens-canis L. and P. officinalis L., Journal of plant physiology, **165**(12), 1255-63

Fasoli. M, Dal Santo. S, Zenoni. S, Tornielli. G. B, Farina. L, Zamboni. A, Porceddu. A, Venturini. L, Bicego. M, Murino. V, Ferrarini. A, Delledonne. M, Pezzotti. M, 2012, The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program, The plant cell, **24**(9), 3489-505

Ferrer. J. L, Jez. J. M, Bowman. M. E, Dixon. R. A, Noel. J. P, 1999, Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis, Nature structural biology, **6**(8), 775-84

Finseth. F. R, Dong. Y, Saunders. A, Fishman. L, 2015, Duplication and adaptive evolution of a key centromeric protein in Mimulus, a genus with female meiotic drive, MBE, **32**(10), 2694-2706

Fishman. L, Kelly. A. J, Willis. J. H, 2002, Minor quantitative trait loci underlie floral traits associated with mating system divergence in Mimulus, Evolution, **56**(11), 2138-2155.

Förster. F, Beisser. D, Grohme. M. A, Liang. C, Mali. B, Siegl. A. M, Engelmann. J. C, Shkumatov. A. V, Schokraie. EMüller. T, Schnölzer. M,Schill. R. O, Frohme. M, Dandekar. T, 2012, Transcriptome analysis in tardigrade species reveals specific molecular pathways for stress adaptations, Bioinformatics and biology insights, **6**, 69-96

Francis. W. R, Christianson. L. M, Kiko. R, Powers. M. L, Shaner. N. C, Haddock. S. H. D, 2013, BMC genomics, **14**(1), 1

Freeling. M, 2009, Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition, Annual review of plant biology, **60**, 433-53

Gao. D, Ko. D. C, Tian. X, Yang. G, Wang. L, 2015, Expression Divergence of Duplicate Genes in the Protein Kinase Superfamily in Pacific Oyster, **11**(1), 57-65

Gayral. P, Melo-Ferreira. J, Glémin. S, Bierne. N, Carneiro. M, Nabholz. B, Lourenco. J. M, Alves. P. C, Ballenghien. M, Faivre. N, Belkhir,. K, Cahais. V, Loire. E, Bernard. A, Galtier. N, 2013, Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap, PloS genetics, **9**(4), e1003457

Gefen. E, Talal. S, Brendzel. O, Dror. A, Fishman, A, 2015, Variation in quantity and composition of cuticular hydrocarbons in the scorpion Buthus occitanus (Buthidae) in response to acute exposure to desiccation stress, Comparative biochemistry and physiology, **182**, 58-63

Geisler. M, Gibson. D. J, Lindsey. K. J, Millar. K, Andrew. J, 2012, Upregulation of photosynthesis genes , and down- regulation of stress defense genes , is the response of Arabidopsis thaliana shoots to intraspecific competition, Botanical studies, **53**, 85-96

Gevers. D, Vandepoele. K, Simillion. C, Van de Peer. Y, 2004, Gene duplication and biased functional retention of paralogs in bacterial genomes, Trends in microbiology, **12**(4), 145-48

Ghanevati. M, Jaworski. J. G, 2001, Active-site residues of a plant membrane-bound fatty acid elongase, Biochimica et biophysica acta, **1530**, 77-85

Gilbert. D, 2013, Gene-omes built from mRNA seq not genome DNA, 7th annual arthropod genomics symposium, Notre Dame

Glover. N. M, Redestig. H, Dessimoz. C, 2016, Homoeologs: What Are They and How Do We Infer Them?, Trends in plant science, **21**(7), 609-21

Goodall-copestake. W. P, Harris. D. J, Hollingsworth. P. M, 2009, The origin of a mega-diverse genus: dating Begonia (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods, Biological Journal of the Linnaean Society, **159**, 363-80

Goodall-copestake. W. P, Pérez-espona. S, Harris. D. J, Hollingsworth. P. M, 2010, The early evolution of the mega-diverse genus Begonia (Begoniaceae) inferred from organelle DNA phylogenies, Biological Journal of the Linnaean Society, **101**, 243-50

Goodstein. D. M, Shengquiang. S, Howson. R, Neupane. R, Hayes. R. D, Fazo, J, Mitros. T, Dirks. W, Hellsten. U, Putnam. N, Rokhsar. D. S, 2012, Phytozome: a comparative platform for green plant genomics, Nucleic Acids Research, **40** (Database Issue), D1178 – D1186

Gould. K. S, McKelvie. J, Markham. K. R, 2002, Do anthocyanins function as antioxidants in leaves? Imaging of H2O2 in red and green leaves after mechanical injury, Plant, cell and environment, **25**(10), 1261-69

Grabherr. M. G, Haas. B. J, Yassour. M, Levin. J. Z, Thompson. D. A, Amit. I, Adiconis. X, Fan. L, Raychowdhury. R, Zeng. Q, Chen. Z, Mauceli. E, Hacohen. N, Gnirke. A, Rhind. N, di Palma. F, Birren. B. W, Nusbaum. C, Lindblad-Toh. K, Friedman. N, Regev. A, 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nature biotechnology, **29**(7), 644-52

Grivet. D, Climent. J, Zabal-Aguirre. M, Neale. N. B, Vendramin. G. G, González-Martínez. S. C, 2013, Adaptive evolution of Mediterranean pines., MPE, **68** (3), 555-56

Gu. Z, Cavalcanti. A, Chen. F, Bouman. P, Li. W, 2002, Extent of gene duplication in the genomes of Drosophila, nematode, and yeast, MBE, **19**(3), 256-62

Gugger. P. F, Cokus. S. J, Sork. V. L, 2016, Association of transcriptome-wide sequence variation with climate gradients in valley oak (Quercus lobata), Tree genetics and genomes, **12**(2), 15

Guo. S, Zheng. Y, Joung. J, Liu. S, Zhang. Z, Crasta. O. R, Sobral. B. W, Xu. Y, Huang. S, Fei. Z, 2010, Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types, BMC genomics, **11**, 384

Hahn. M. W, Han. M. V, Han. S, 2007, Gene family evolution across 12 Drosophila genomes, PloS genetics, **3**(11), e197

Han. Y, Zhao. W, Wang. Z, Zhu. J, Liu. Q, 2014, Molecular evolution and sequence divergence of plant chalcone synthase and chalcone synthase-Like genes, Genetica, **142**(3), 215-25

Hanso. M, Drenkhan. R, 2008, First observations of *Mycospaerella pini* in Estonia, Plant pathology, **57**, 1177

Hanks. M, Wurst. W, Anson-Cartwright. L, Auerbach. A. B, Joyner. A. L, 1995, Rescue of the En-1 Mutant Phenotype by Replacement of En-1 with En-2, Science, **269**(5224), 679

Harikrishnan. S. L, Pucholt. P, Berlin. S, 2015, Sequence and gene expression evolution of paralogous genes in willows, Scientific reports, **5**(11), 18662

Harris. R, Hofmann. H, 2015, Seeing is believing: Dynamic evolution of gene families, PNAS, **112**(5), 1252 – 1253

Hartmann. S, Helm. C, Nickel. B, Meyer. M, Struck. T. H, Tiedemann. R, Selbig. J, Bleidorn. C, 2012, Exploiting gene families for phylogenomic analysis of myzostomid transcriptome data, PloS one, **7**(1), e29843

He. X, Zhang. J, 2005, Gene complexity and gene duplicability, Current biology, **15**(11), 1016-21

Heim. K. E, Tagliaferro. A. R, Bobilya. D. J, 2002, Flavonoid antioxidants: chemistry, metabolism and structure-activity relationships, The journal of nutritional biochemistry, **13**(10), 572-84

Hemmer-Hansen. J, Therkildsen. N. O, Meldrup. D, Nielsen. E. E, 2013, Conserving marine biodiversity: insights from life-history trait candidate genes in Atlantic cod (Gadus morhua), Conservation genetics, **15**(1), 213-228

Hoegg. Simone, Brinkmann. H, Taylor. J. S, Meyer. A, 2004, Phylogenetic Timing of the Fish-Specific Genome Duplication Correlates with the Diversification of Teleost Fish, JME, **59**, 190-203

Huang. S, Li. R, Zhang. Z, Li. L, Gu. X, Fan. W, Lucas. W. J, Wang. X, Xie. B, Ni. P, Ren. Y, Zhu. H, Li. J, Lin. K, Jin. W, Fei. Z, Li. G, Staub. J, Kilian. A, van der Vossen. E. A. G, Wu. Y, Guo. J, He. J, Jia. Z, Ren. Y, Tian. G, Lu. Y, Ruan. J, Qian. W, Wang. M, Huang. Q, Li. B, Xuan. Z, Cao. J. A, Wu. Z, Zhang. J, Cai. Q, Bai. Y, Zhao. B, Han. Y, Li. Y, Li. X, Wang. S, Shi. Q, Liu. S, Cho. J, Kim. J, Xu. Y, Heller-Uszynska. K, Miao. H, Cheng. Z, Zhang. S, Wu. J, Yang. Y, Kang. H, Li. M, Liang. H, Ren. X, Shi. Z, Wen. M, Jian. M, Yang. H, Zhang. G, Yang. Z, Chen. R, Liu. S, Li. J, Ma. J, Liu. H, Zhou. Y, Zhao. J, Fang. X, Li. G, Fang. L, Li. Y, Liu. D, Zheng. H, Zhang. Y, Qin. N, Li. Z, Yang. G, Yang. Z, Bolund. L, Kristiansen. K, Zheng. H, Li. S, Zhang. X, Yang. H, Wang. J, Sun. R, Zhang. B, Jiang. S, Wang. J, Du. Y, Li. S, 2009, The genome of the cucumber, Cucumis sativus L., Nature genetics, **41**(12), 1275-81

Huang. B, Pang. E, Chen. Y, Cao. H, Ruan. Y, Liao. P, 2015, Positive selection and functional divergence of R2R3-MYB paralogous genes expressed in inflorescence buds of Scutellaria species (Labiatae), International journal of molecular sciences, **16**(3), 5900-21

Huang. X, Chen. X, Armbruster. P. A, 2016, Comparative performance of transcriptome assembly methods for non-model organisms, BMC genomics, **17**, 523

Hughes. M, Hollingsworth. P. M, Squirrel. J, 2002, Isolation of polymorphic microsatellite markers for Begonia sutherlandii Hook. f., Molecular ecology notes, **2**, 185-6

Hughes. M, Hollingsworth. P. M, Miller. A. G, 2003, Population genetic structure in the endemic Begonia of the Socotra archipelago, Biological conservation, **113**(2), 277-84

Hughes. M, Hollingsworth. P. M, 2008, Population genetic divergence corresponds with species-level biodiversity patterns in the large genus Begonia., Molecular ecology, **17**(11), 2643-51

Hughes. N. M, Vogelmann. T. C, Smith. W. K, 2008, Optical effects of abaxial anthocyanin on absorption of red wavelengths by understorey species: revisiting the back-scatter hypothesis, Journal of experimental botany, **59**(12), 3435-42

Hughes. N. M, Carpenter. K. L, Keidel. T. S, Miller. C. N, Waters. M. N, Smith. W. K, 2014, Photosynthetic costs and benefits of abaxial versus adaxial anthocyanins in Colocasia esculenta 'Mojito', Planta, **240**(5), 971-81

Jiao. Y, Wickett. N. J, Ayyampalayam. S, Chanderbali. A. S, Landherr. L, Ralph. P. E, Tomsho. L. P, Hu. Y, Liang. H, Soltis. P. S, Soltis. D. E, Clifton. S. W, Schlarbaum. S. E, Schuster. S. C, Ma. H, Leebens-Mack. JDePamphilis. C. W, 2011, Ancestral polyploidy in seed plants and angiosperms, Nature, **473**(7345), 97-100

Johnson. A. D, Handsaker. R. E, Pulit. S. L, Nizzari. M. M, O'Donnell. C. J, de Bakker. P. I. W, 2008, **24**(24), 2938-9

Kachi. N, Okuda. T, Numata. S, Manokaran. N, 2004, Delayed greening, leaf expansion, and damage to sympatric Shorea species in a lowland rain forest, Journal of plant research, **117**(1), 19-25

Kannan. S, Hui. J, Mazooji. K, 2016, Shannon: An Information-Optimal de Novo RNA-Seq Assembler, bioRxiv

Karageorgou. P, Buschmann. C, Manetas. Y, 2008, Red leaf color as a warning signal against insect herbivory: Honest or mimetic?, Flora, **203**(8), 648-52

Katoh. K, Misawa. K, Kuma. K, Miyata. T, 2002, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic acids research, **30**(14), 3049-66

Kearse. M, Moir. R, Wilson. A, Stones-Havas. S, Cheung. M, Sturrock. S, Buxton. S, Cooper. A, Markowitz. S, Duran. C, Thierer. T, Ashton. B, Meintjes. P, Drummond. A, 2012, Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics, **28**(12), 1647-9

Koch. M. A, Haubold. B, Mitchell-Olds. T, 2000, Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in Arabidopsis, Arabis, and Related Genera (Brassicaceae), MBE, **17**(10), 1483-98

Koenig. D, Weigel. D, 2015, Beyond the thale: comparative genomics and genetics of Arabidopsis relatives, Nature reviews genetics, **16**(5), 285-98

Kondrashov. F. A, Rogozin. I. B, Wolf. Y. I, Koonin. E. V, 2002, Selection in the evolution of gene duplications, Genome biology, **3**(2), 8

Kong. J, Chia. L, Goh. N, Chia. T, Brouillard. R, 2003, Analysis and biological activities of anthocyanins, Phytochemistry, **64**(5), 923-33

Kosma. D. K, Bourdenx. B, Bernard. A, Parsons. E. P, Lü. S, Joubès. J, Jenks. M. A, 2009, The impact of water deficiency on leaf cuticle lipids of Arabidopsis, Plant physiology, **151**(4), 1918-29

Koutsovoulos. G. D, 2015, Reconstructing the phylogenetic relationships of nematodes using draft genomes and transcriptomes, PhD Thesis

Krause. G. H, Winter. K, Matsubara. S, Krause. B, Jahns. P, Virgo. A, Aranda. J, García. M, 2012, Photosynthesis, photoprotection, and growth of shade-tolerant tropical tree seedlings under full sunlight, Photosynthesis research, **113**(1), 273-85

Künstner. A,Wolf. J. B. W, Backström. N, Whitney. O, Balakrishnan. C. N, Day. L, Edwards. S. V, Janes. D. E, Schlinger. B. A, Wilson. R. K, Jarvis. E. D, Warren. W. C, Ellegren. H, 2010, Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species, Molecular ecology, **19**, 266-76

Lamesch. P, Berardini. T. Z, Li. D, Swarbreck. D, Wilks. C, Sasidharan. R, Muller. R, Dreher. K, Alexander. D. L, Garcia-Hernandez. M, Karthikeyan. A. S, Lee. C. H, Nelson. W. D, Ploetz. L, Singh. S, Wensel. A, Huala. E, 2012, The Arabidopsis Information Resource (TAIR)L improved gene annotation and new tools, Nucleic Acids Research, **40**, Database Issue D1202-10

Laukaitis. C. M, Heger. A, Blakley. T. D, Munclinger. P, Ponting. C. P, Karn. R. C, 2008, Rapid bursts of androgen-binding protein (Abp) gene duplication occurred independently in diverse mammals, BMC evolutionary biology, **8**, 46

Lee. D. W, Lowry. J. B, Stone. B. C, 1979, Abaxial anthocyanin layer in leaves of tropical rain forest plants: enhancer of light capture in deep shade, Biotropica, **11**(1), 70-77

Lee. S, Jung. S, Go. Y, Kim. H, Kim. J, Cho. H, Park. O, Suh. M, 2009, Two Arabidopsis 3-ketoacyl CoA synthase genes, KCS20 and KCS2/DAISY, are functionally redundant in cuticular wax and root suberin biosynthesis, but differentially controlled by osmotic stress, The plant journal, **60**(3), 462-75

Leitch. A, Leitch. I, 2008, Genomic plasticity and the diversity of polyploid plants, Science, **320**(5875), 481-3

Lev-Yadun. S, Dafni. A, Flaishman. M. A, Inbar. M, Izhaki. I, Katzir. G, Ne'eman. G, 2004, Plant coloration undermines herbivorous insect camouflage, Bioessays, **26**(10), 1126-30

Lev-yadun. S, Gould. K. S, 2009, Role of Anthocyanins in Plant Defence, In Anthocyanins, 22-28, New York, Springer

Li. X, Noll. M, 1994, Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions, Nature, **367**(6458), 83-87

Lohse. M, Nagel. A, Herter. T, May. P, Schroda, M, Zrenner. R, Tohge. T, Fernie. A. R, Stitt. M, Usadel. B, 2014, Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data, Plant, cell & environment, 37(5), 1250-1258

Liu. Y, Schroder. J, Schmidt. B, 2013, Musket: A multistage k-mer spectrum-based error corrector for Illumina sequence data, Bioinformatics, **29**(3), 308-15

Long. S. P, Humphries. S, 1994, Photoinhibition of photosynthesis in nature, Annual reviews of plant physiology and plant molecular biology, **45**, 633-62

Lu. J, Peatman. E, Tang. H, Lewis. J, Liu. Z, 2012, Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications, BMC genomics, **13**, 246

Luo. R, Liu. B, Xie. Y, Li. Z, Huang. W, Yuan. J, He. G, Chen. Y, Pan. Q, Liu. Y, Tang. J, Wu. G, Zhang. H, Shi. Y, Liu. Y, Yu. C, Wang. B, Lu. Y, Han. C, Cheung. D. W, Yiu. S, Peng. S, Xiaoqian. Z, Liu. G, Liao. X, Li. Y, Yang. H, Wang. J, Lam. T, Wang. J, 2012, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience, **1**(1), 18

Lunt. D. H, Kumar. S, Koutsovoulos. G, Blaxter. M. L, 2014, The complex hybrid origins of the root knot nematodes revealed through comparative genomics, PeerJ, **2**, e356

Lynch. M, Conery. J. S, 2000, The Evolutionary Fate and Consequences of Duplicate Genes, Science, **290**(5494), 1151-1155

Lynch. M, Katju. V, 2004, The altered evolutionary trajectories of gene duplicates, Trends in genetics, **20**(11), 544-9

Lyons. E, Freeling. M, 2008, How to usefully compare homologous plant genes and chromosomes as DNA sequences, The Plant Journal, **53**(4), 661-673

Ma. X, Wang. P, Zhou. S, Sun. Y, Liu. N, Li. X, Hou. Y, 2015, De novo transcriptome sequencing and comprehensive analysis of the drought-responsive genes in the desert plant Cynanchum komarovii, **16**(1), 753

Mastretta-Yanes. A, Arrigo. N, Alvarez. N, Jorgensen. T. H, Piñero. D, Emerson. B. C, 2014, Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference, Molecular ecology resources, **15**(1), 28-41

Moghadam. H. K, Harrison. P. W, Zachar. G, Székely. T, Mank. J. E, 2013, The plover neurotranscriptome assembly: transcriptomic analysis in an ecological model species without a reference genome, Molecular ecology resources, **13**(4), 696-705

Mendoza. C, Naumann. J, Samain. M, Goetghebeur. P, De Smet. Y, Wanke. S, 2015, A genome-scale mining strategy for recovering novel rapidly-evolving nuclear single-copy genes for addressing shallow-scale phylogenetics in Hydrangea, BMC evolutionary biology, **15**(1), 132

Macmanes. M. D, Eisen. M. B, 2013, Improving transcriptome assembly through error correction of high-throughput sequence reads, PeerJ, **1**, e113

Manetas. Y, Drinia. A, Petropoulou. Y, 2002, High contents of anthocyanins in young leaves are correlated with low pools of xanthophyll cycle components and low risk of photoinhibition, Photosynthetica, **40**(3), 349-54

Manetas. Y, 2006, Why some leaves are anthocyanic and why most anthocyanic leaves are red?, Flora, **201**(3), 163-77

McClintock. B, 1983, The significance of responses of the genome to challenge, Nobel lecture

McCormack. J. E, Harvey. M. G, Faircloth. B. C, Crawfod. N. G, Glenn. T. C, Brumfield. R. T, 2013, A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing, PloS one, **8** (1), e54848

Meunier. J, Lemoine. F, Soumillon. M, Liechti. A, Weier. M, Guschanski. K, Hu. H, Khaitovich. P, Kaessmann. H, 2013, Birth and expression evolution of mammalian microRNA genes, Genome research, **23**(1), 34-45

Monson. R. K, 2003, Gene Duplication, Neofunctionalization, and the Evolution of C4 Photosynthesis, International journal of plant sciences, **164**(3), 43-54

Moonlight. P. W, Richardson. J. E, Tebbitt. M. C, Thomas. D. C, Hollands. R, Peng. C, Hughes. M, 2015, Continental-scale diversification patterns in a megadiverse genus: the biogeography of Neotropical Begonia, Journal of biogeography, **46**(2), 1137-1149

Moore. R, Purugganan. M, 2003, The early stages of duplicate gene evolution, PNAS, **100**(26), 15682-7

Nagalakshmi. U, Wang. Z, Waern. K, Shou. C, Raha. D, Gerstein. M, Snyder. M, 2008, The transcriptional landscape of the yeast genome defined by RNA sequencing, Science, **320**(5881), 1344-1349.

Nakasugi. K, Crowhurst. R, Bally. J, Waterhouse. P, 2014, Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana, PloS one, **9**(3), e91776

Neale. S, Goodall-Copestake. W, Kidner. C. A, Teixeira da Silva. J, 2006, The evolution of diversity in *Begonia*, Floriculture, Ornamental and Plant Biotechnology, 608345797

Neill. S. O, Gould. K. S, Kilmartin. P. A, Mitchell. K. A, Markham. K. R, 2002, Antioxidant activities of red versus green leaves in Elatostema rugosum, Plant, cell and environment, **25**(4), 539-47

Nikolenko. S. I, Korobeynikov. A. I, Alekseyev. M. A, 2013, BayesHammer: Bayesian clustering for error correction in single-cell sequencing, BMC genomics, **14**(1), 7

Oakley. T. H, Wolfe. J. M, Lindgren. A. R, Zaharoff. A. K, 2013, Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny, MBE, **30**(1), 215-33

Ohno, S, (2013) *Evolution by gene duplication*, Springer Science & Business Media

Oppenheim. S. J, Baker. R. H, Simon. S, DeSalle. R, 2014, We can't all be supermodels: the value of comparative transcriptomics to the study of non-model insects, Insect molecular biology, **24**(2), 139-154

Pál. C, Papp. B, Hurst, L. D, 2001, Does the Recombination Rate Affect the Efficiency of Purifying Selection? The Yeast Genome Provides a Partial Answer, MBE, **18**(12), 2323-26

Patel. R. K, Jain. M, 2012, NGS QC Toolkit: a toolkit for quality control of next generation sequencing data, PloS one, **7**(2), e30619

Paterson. A. H, Chapman. B. A, Kissinger. J. C, Bowers. J. E, Feltus. F. A, Estill. J. C, 2006, Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon, Trends in genetics, **22**(11), 597-602

Patro. Rob, Duggal. Geet, Love. M. I, Irizarry. R. A, Kingsford. C, 2015, Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment, bioRxiv

Paul. S, Gable. K, Beaudoin. F, Cahoon. E, Jaworski. J, Napier. J. A, Dunn. T. M, 2006, Members of the Arabidopsis FAE1-like 3-ketoacyl-CoA synthase gene family substitute for the Elop proteins of Saccharomyces cerevisiae, Journal of biological chemistry, **281**(14), 9018-29

Pease. J. B, Haak. D. C, Hahn. M. W, Moyle. L. C, 2016, Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation, PloS biology, **14**(2), e1002379

Pegueroles. C, Laurie. S, Albà. M. M, 2013, Accelerated evolution after gene duplication: a time-dependent process affecting just one copy, MBE, **30**(8), 1830-42

Petrov. D, 2001, Evolution of genome size: new approaches to an old problem, Trends in genetics, **17**(1), 23-28

Plana. V, Gascoigne. A, Forrest. L. L, Harris. D, Pennington. T. R, 2004, Pleistocene and pre-Pleistocene Begonia speciation in Africa., MPE, **31**(2), 449-61

Posada. D, 2008, jModelTest: phylogenetic model averaging, MBE, **25**(7), 1253-6

Postnikova. O. A, Hult. M, Shao. J, Skantar. A, Nemchinov. L. G, 2015, Transcriptome analysis of resistant and susceptible alfalfa cultivars infected with root-knot nematode Meloidogyne incognita, PloS one, **10**(2), 1-17

Puttick. M. N, Clark. J, Donoghue. P. C. J, 2015, Size is not everything : rates of genome size evolution , not C -value , correlate with speciation in angiosperms, Proceedinds of the Royal Society B, **282**(1820)

Quail. M. A, Smith. M, Coupland. P, Otto. T. D, Harris. S. R, Connor. T. R, Bertoni. A, Swerdlow. H. P, Gu. Y, 2012, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC genomics, **13**(1), 341

Quince. C, Lanzén. A, Curtis. T. P, Davenport. R. J, Hall. N, Head. I. M, Read. L. F, Sloan. W. T, 2009, Accurate determination of microbial diversity from 454 pyrosequencing data, Nature methods, **6**(9), 639-41

Ranwez. V, Harispe. S, Delsuc. F, Douzery. E. J. P, 2011, MACSE: Multiple Alignment of Coding Sequences accounting for frameshifts and stop codons, PloS one, **6**(9), e22594

Renny-Byfield. S, Wendel. J. F, 2014, Doubling down on genomes: polyploidy and crop plants, American Journal of Botany, 101(**10**), 1711-1725

Rice. P, 2000, The European Molecular Biology Open Software Suite EMBOSS: The European Molecular Biology Open Software Suite, Trends in genetics, **16**(6), 2-3

Richard. S, Lapointe. G, Rutledge. R. G, Séguin. A, 2000, Induction of Chalcone Synthase Expression in White Spruce by Wounding and Jasmonate, Plant cell physiology, **41**(8), 982-87

Riederer. M, Schreiber. L, 2001, Protecting against water loss: analysis of the barrier properties of plant cuticles., Journal of experimental botany, **52**(363), 2023-32

Robert. C, Watson. M, 2015, Errors in RNA-Seq quantification affect genes of relevance to human disease, Genome biology, **16**(1), 177

Robertson. G, Schein. J, Chiu. R, Corbett. R, Field. M, Jackman. S. D, Mungall. K, Lee. S, Okada. H. M, Qian. J. Q, Griffith. M, Raymond. A, Thiessen. N, Cezard. T, Butterfield. Y. S, Newsome. R, Chan. S. K, She. R, Varhol. R, Kamoh. B, Prabhu. A, Tam. A, Zhao. Y, Moore. R. A, Hirst. M, Marra. M. A, Jones. S. J. M, Hoodless. P. A, Birol. I, 2010, De novo assembly and analysis of RNA-seq data, Nature methods, **7**(11), 909-12

Robinson. M. D, McCarthy. D. J, Smyth. G. K, 2010, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, **26**(1), 139-40

Ronquist, F, Huelsenbeck. J. P, 2003, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics, **19**(12), 1572-74

Rothfels. C. J, Larsson. A, Li. F, Sigel. E. M, Huiet. L, Burge. D. O, Ruhsam. M, Graham. S. W, Stevenson. D. W, Wong. G. K, Korall. P, Pryer. K. M, 2013, Transcriptome-mining for single-copy nuclear markers in ferns, PloS one, **8**(10), e76957

Rozen. S, Skaletsky. H, 1999, Primer3 on the WWW for general users and for biologist programmers, Bioinformatics methods and protocols, 365-386

Sakuma. S, Pourkheirandish. M, Hensel. G, Kumlehn. J, Stein. N, Tagiri. A, Yamaji. A, Ma. J. F, Sassa. H, Koba. T, Komatsuda. T, 2013, Divergence of expression pattern contributed to neofunctionalization of duplicated HD-Zip I transcription factor in barley, New phytologist, **197**(3), 939-48

Salmela. L, Rivals. E, 2014, LoRDEC: Accurate and efficient long read error correction, Bioinformatics, **30**(24), 3506-14

Schaefer. M. H, Rolshausen. G, 2006, Plants on red alert: do insects pay attention?, BioEssays, **28**(1), 65-71

Schmutz. J, Cannon. S. B, Schlueter. J, Ma. J, Mitros. T, Nelson. W, Hyten. D. L, Song. Q, Thelen. J. J, Cheng. J, Xu. D, Hellsten. U, May. G. D, Yu. Y, Sakurai. T, Umezawa. T,

Bhattacharyya. M. K, Sandhu. D, Valliyodan. B, Lindquist. E, Peto. M, Grant. D, Shu. S, Goodstein. D, Barry. K, Futrell-Griggs. M, Abernathy. B, Du. J, Tian. Z, Zhu. L, Gill. N, Joshi. T, Libault. M, Sethuraman. A, Zhang. X. C, Shinozaki. K, Nguyen. H. T, Wing. R. A, Cregan. P, Specht. J, Grimwood. J, Rokhsar. D, Stacey. G, Shoemaker. R. C, Jackson. S. A, 2010, Genome sequence of the palaeopolyploid soybean, Nature, **463**(7278), 178-83

Schnable. J. C, Springer. N. M, Freeling. M, 2011, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss, PNAS, **108**(10), 4069-74

Schmieder. R, Edwards. R, 2011, Quality control and preprocessing of metagenomic datasets, Bioinformatics, **27**(6), 863-64

Schröder. J, 1997, A family of plant-specific polyketide synthases: facts and predictions, Trends in plant science, **2**(10), 373-378

Schulz. E, Tohge. T, Zuther. E, Fernie. A. R, Hincha. D. K, 2015, Natural variation in flavonol and anthocyanin metabolism during cold acclimation in Arabidopsis thaliana accessions, Plant, cell and environment, **38**(8), 1658-72

Sedeek. K. E. M, Qi. W, Schauer. M. A, Gupta. A. K, Poveda. L, Xu. S, Liu. Z, Grossniklaus. U, Schiestl. F. P, Schlüter. P. M, 2013, Transcriptome and proteome data reveal candidate genes for pollinator attraction in sexually deceptive orchids, PloS one, **8**(5), e64621

Seo. P, Park. C, 2011, Cuticular wax biosynthesis as a way of inducing drought resistance, Plant signalling and behaviour, **6**(7), 1043-45

Serrano. M, Coluccia. F, MarthaTorres, F. L. H, Métraux. J. P, 2015, The cuticle and plant defense to pathogens, Plant cell wall in pathogenesis, parasitism and symbiosis, 6

Sievers. F, and Higgins. D. G, 2014, Clustal Omega, accurate alignment of very large numbers of sequences, Multiple sequence alignment methods, 105-116

Simão. F. A, Waterhouse. R. M, Ioannidis. P, Kriventseva. E. V, Zdobnov. E. M, 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, doi: 10.1093/bioinformatics/btv351

Simillion. C, Vandepoele. K, Van Montagu. M. C. E, Zabeau. M, Van de Peer. Y, 2002, The hidden duplication past of Arabidopsis thaliana, PNAS, **99**(21), 13672-32

Slotte. T, Hazzouri. K. M, Stern. D, Andolfatto. P, Wright. S. I, 2012, Genetic architecture and adaptive significance of the selfing syndrome in Capsella, Evolution, **66**(5), 1360-74

Small. R. L, Wendel. J. F, 2000, Phylogeny, duplication, and intraspecific variation of Adh sequences in New World diploid cottons (Gossypium l., malvaceae), Molecular phylogenetics and evolution, **16**(1), 73-84

Smillie. R. M, Hetherington. S. E, 1999, Photoabatement by anthocyanin shields photosynthetic systems from light stress, Photosynthetica, **36**(3), 451-63

Smit. A. F. A, & Hubley. R, 2010, RepeatModeler Open-1.0. http://www.repeatmasker.org/

Smith-Unna. R, Boursnell. C, Patro. R, Hibberd. J, Kelly. S, 2016, TransRate: reference free quality assessment of de novo transcriptome assemblies, *Genome research*, gr-196469.

Soltis. D. E, Buggs. R. J. A, Barbazuk. W. B, Chester. M, Gallagher. J. P, Schnable. P. S, Soltis. P. S, 2012, The Early Stages of Polyploidy: Rapid and Repeated Evolution in Tragopogon in, Soltis. D. E, Polyploidy and Genome Evolution (271 – 292), Heidelberg: Springer-Verlag

Stamatakis. A, 2014, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics, **30**(9), 1312-13

Stanke. M, Morgenstern. B, 2005, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic acids research, **33**, 465-7

Steyn. W. J, Wand. S. J. E, Holcroft. D. M, Jacobs. G, 2002, Anthocyanins in vegetative tissues: a proposed unified function in photoprotection, New phytologist, **155**(3), 349-61

Suyama. M, Torrents. D, Bork. P, 2006, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, Nucleic acids research, **34**, 609-12

Sveinsson. S, McDill. J, Wong. G. K. S, Li. J, Li. X, Deyholos. M. K, Cronk. Q. C. B, 2014, Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (Linum) using transcriptomics, Annals of botany, **113**(5), 753-61

Takos. A. M, Jaffé. F. W, Jacob. S. R, Bogs. J, Robinson. S. P, Walker. A. R, 2006, Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples, Plant physiology, **142**(3), 1216-32

Tang. F, Lao. K, Surani. M. A, 2011, Development and applications of single-cell transcriptome analysis, Nature methods, **8**(4), 6-11

Tate. J. A, Symonds. V. V, Doust. A. N, Buggs. R. J. A, Mavrodiev. E, Majure. L. C, Soltis. P. S, Soltis. D. E, 2009, Synthetic polyploids of Tragopogon miscellus and T. mirus (Asteraceae): 60 Years after Ownbey's discovery, American journal of botany, **96**(5), 979-88

Thalmann. M. R, Pazmino. D, Seung. D, Horrer. D, Nigro. A, Meier. T, Kölling. K, Pfeifhofer. H. W, Zeeman. S. C, Santelia. D, 2016, Regulation of leaf starch degradation by abscisic acid is important for osmotic stress tolerance in plants, The plant cell, tpc-00143

Throude. M, Bolot. S, Bosio. M, Pont. C, Sarda. X, Quraishi. U. M, Bourgis. F, Lessard. P, Rogowsky. P, Ghesquiere. A, Murigneux. A, Charmet. G, Perez. P, Salse. J, 2009, Structure and expression analysis of rice paleo duplications, Nucleic acids research, **37**(4), 1248-59

Todaka. D, Shinozaki, K, Yamaguchi-Shinozaki. K, 2015, Recent advances in the dissection of drought-stress regulatory networks and strategies for development of drought-tolerant transgenic rice plants, Frontiers in plant science, **6**(84)

Todd. J, Post-beittenmiller. D, Jaworski. J. G, 1999, KCS1 encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in Arabidopsis thaliana, The plant journal, **17**, 119-30

Tropf. S, Kärcher. B, Schröder. G, Schröder. J, 1995, Reaction mechanisms of homodimeric plant polyketide synthases (Stilbene and Chalcone synthase), The journal of biological chemistry, **14**(4), 7922-28

Tsagkogeorga. G, Turon. X, Galtier. N, Douzery. E. J. P. Delsuc. F, 2010, Accelerated evolutionary rate of housekeeping genes in tunicates, JME, **71** (2), 153-67

Twyford. A. D, Kidner. C. A, Harrison. N, Ennos. R. A, 2013, Population history and seed dispersal in widespread Central American Begonia species (Begoniaceae) inferred from plastome-derived microsatellite markers, Botanical journal of the Linnaean Society, **171**, 260-76

Twyford. A. D, Kidner. C. A, Ennos. R. A, 2015, Maintenance of species boundaries in a Neotropical radiation of Begonia., Molecular ecology, **24**(19), 4982-93

Van Megen. H, Ven Den Elsen. S, Holterman. M, Karssen. G, Mooyman. P, Bongers. T, Holovachov. O, Bakker. J, Helder. J, 2009, A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences, Nematology, **11**(6), 927-50

Verdier. J, Torres-Jerez. I, Wang. M, Andriankaja. A, Allen. S. N, He. J, Tang. Y, Murray. J. D, Udvardi. M. K, 2013, Establishment of the Lotus japonicus Gene Expression Atlas (LjGEA) and its use to explore legume seed maturation, The plant journal, **74**(2), 351-62

Vijay. Nagarjun, Poelstra. J. W, Künstner. A, Wolf. J. B. W, 2013, Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments, Molecular ecology, **22**(3), 620-34

Visser. E. A, Wegrzyn. J. L, Steenkmap. E. T, Myburg. A. A, Naidoo. S, 2015, Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome, BMC genomics, **16**(1), 1057

Wagner. A, 2001, Birth and death of duplicated genes in completely sequenced eukaryotes., Trends in genetics, **17**(5), 237-9

Wang. Z, Gerstein. M, Snyder. M, 2009, RNA-Seq: a revolutionary tool for transcriptomics., Nature reviews genetics, **10**(1), 57-63

Wang. J, Tao. F, Marowsky. N. C, Fan. C, 2016, Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes, Plant physiology, 1177

Wei. W, Qi. X, Wang. L, Zhang. Y,Hua. W, Li. D, Lv. H, Zhang, Xiurong, 2011, Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers, BMC genomics, **12**(1), 451

Weidenbach. D, Jansen. M, Franke. R. B, Hensel. G, Weissgerber. W, Ulferts. S, Jansen. I, Schreiber. L, Korzun. V, Pontzen. R, Kumlehn. J, Pillen. K, Schaffrath. U, 2014, Evolutionary Conserved Function of Barley and Arabidopsis 3-KETOACYL-CoA SYNTHASES in Providing Wax Signals for Germination of Powdery Mildew Fungi, Plant physiology, **166**(3), 1621-33

Weiss-Schneeweiss. H, Blöch. C, Turner. B, Villaseñor. J. L, Stuessy. T. F, Schneeweiss. G. M, 2012, The promiscuous and the chaste: frequent allopolyploid speciation and its genomic consequences in American daisies (Melampodium sect. Melampodium; Asteraceae), Evolution, **66**(1), 211-228

Wernersson. R, 2003, RevTrans: multiple alignment of coding DNA from aligned amino acid sequences, Nucleic acids research, **31**(13), 3537-39

Yampolsky. L. Y, Stoltzfus. A, 2005, The exchangeability of amino acids in proteins, Genetics, **170**(4), 1459-72

Yang. J, Huang. J, Gu, H, Zhong. Y, Yang. Z, 2002, Duplication and Adaptive Evolution of the Chalcone Synthase Genes of Dendranthema (Asteraceae), MBE, **19**(10), 1752-59

Yang. J, Gu. H, Yang. Z, 2004, Likelihood analysis of the chalcone synthase genes suggests the role of positive selection in morning glories (Ipomoea), Journal of molecular evolution, **58**(1), 54-63

Yang. Y, Wang. Y, Gao. Y, Zhou. Y, Zhang. E, Hu. Y, Yuan. Y, Liang. G, Xu. C, 2014, Adaptive evolution and divergent expression of heat stress transcription factors in grasses, BMC evolutionary biology, **14**, 147

Yang. Y, Moore. M. J, Brockington. S. F, Soltis. D. E, Wong. G. K, Carpenter. E. J, Zhang. Y, Chen. L, Yan. Z, Xie. Y, Sage. R. F, Covshoff. S, Hibberd. J. M, Nelson. M. N, Smith. S. A, 2015, Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing, MBE, **32**(8), 2001-14

Yi. W. C, Lee. B, Jang. C. S, 2009, Expression diversity and evolutionary dynamics of rice duplicate genes, Molecular genetics and genomics, **281**(5), 483-93

Yuan. Y, Rebocho. A. B, Sagawa. J. M, Stanley. L. E, Bradshaw. H. D, 2016, Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species, PNAS, **113**(9), 2448-53

Zavala. K, Opazo. J. C, 2015, Lineage-Specific Expansion of the Chalcone Synthase Gene Family in Rosids, PloS one, **10**(7), e0133400

Zhang. J, Broeckling. C. D, Blancaflor. E. B, Sledge. M. K, Sumner. L. W, Wang. Z, 2005, Overexpression of WXP1, a putative Medicago truncatula AP2 domain-containing

transcription factor gene, increases cuticular wax accumulation and enhances drought tolerance in transgenic alfalfa (Medicago sativa), The plant journal, **42**(5), 689-707

Zhang. G, Li. B, Li. C, Gilbert. M. T. P, Jarvis. E. D, Wang. J, 2014, Comparative genomic data of the Avian Phylogenomics Project, GigaScience, **3**(1), 26

Zhang. K. M, Wang. J. W, Guo. M. L, Du. W. L, Wu. R. H, Wang. X, 2016, Short-day signals are crucial for the induction of anthocyanin biosynthesis in Begonia semperflorens under low temperature condition, Journal of plant physiology, **204**(1), 1-7

Zhao. Q, Wang. Y, Kong. Y, Luo. D, Li. X, Hao. P, 2011, Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study, BMC bioinformatics, **12**(14), 2

Zhao. Shanrong, Zhang. Baohong, 2015, A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification., BMC genomics, **16**(1), 1

Zuellig. M. P, Kenney. A. M, Sweigart. A. L, 2014, Evolutionary genetics of plant adaptation: insights from new model systems, Current opinion in plant biology, **18**, 44-50

# Appendices

| | CON ff | CON leaf | CON mf | CON pet | CON root | CON vb |
|---|---|---|---|---|---|---|
| CON ff | | | | | | |
| CON leaf | 963 2738 849 | | | | | |
| CON mf | 944 2707 899 | 1172 2060 1318 | | | | |
| CON pet | 817 2930 803 | 850 2825 875 | 1191 2254 1105 | | | |
| CON root | 850 2499 1201 | 997 2419 1134 | 1254 1889 1407 | 1064 2255 1231 | | |
| CON vb | 759 2908 883 | 767 2864 919 | 1290 1972 1288 | 870 2741 939 | 999 2879 672 | |

Appendix 1. Up and down regulated genes between pairwise comparisons of *B. conchifolia* tissues. Numbers denote differential regulation in tissue on top relative to tissue on left; numbers in blue indicate number of genes that are downregulated, in black no change, and in red upregulated.

| | PLE ff | PLE leaf | PLE mf | PLE pet | PLE root | PLE vb |
|---|---|---|---|---|---|---|
| PLE ff | | | | | | |
| PLE leaf | 985 2749 816 | | | | | |
| PLE mf | 807 2925 818 | 1178 2185 1187 | | | | |
| PLE pet | 623 3303 624 | 561 3253 736 | 963 2710 877 | | | |
| PLE root | 956 2291 1303 | 1330 1942 1278 | 1222 1945 1383 | 895 2477 1178 | | |
| PLE vb | 742 2931 877 | 814 2797 939 | 1147 2273 1130 | 505 3494 551 | 1063 2748 739 | |

Appendix 2. Up and down regulated genes between pairwise comparisons of *B. plebeja* tissues. Numbers denote differential regulation in tissue on top relative to tissue on left; numbers in blue indicate number of genes that are downregulated, in black no change, and in red upregulated.

| | CON ff | CON leaf | CON mf | CON pet | CON root | CON vb |
|---|---|---|---|---|---|---|
| PLE ff | 527 3435 588 | 981 2493 1076 | 1227 2186 1137 | 1089 2473 988 | 1267 2308 975 | 1060 2523 967 |
| PLE leaf | 1156 2282 1112 | 552 3482 516 | 1417 1776 1357 | 1073 2471 1006 | 1262 2010 1278 | 1096 2401 1053 |
| PLE mf | 993 2510 1047 | 1233 2055 1262 | 695 3102 753 | 1133 2212 1205 | 1396 1919 1235 | 1215 2078 1257 |
| PLE pet | 807 2934 809 | 804 2851 895 | 1219 2195 1136 | 567 3460 523 | 1211 2394 945 | 867 2882 801 |
| PLE root | 1106 2029 1415 | 1256 2012 1282 | 1374 1638 1538 | 1214 1921 1415 | 232 4050 268 | 887 2442 1221 |
| PLE vb | 911 2593 1046 | 991 2477 1082 | 1354 1842 1354 | 991 2527 1032 | 1062 2644 844 | 502 3533 515 |

Appendix 3. Up and down regulated genes between pairwise comparisons of *B. conchifolia* and *B. plebeja* tissues. Numbers denote differential regulation in tissue on top relative to tissue on left; numbers in blue indicate number of genes that are downregulated, in black no change, and in red upregulated.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf05713g0001.1 | GO:0006952 | defense response |
| Becon104Scf05956g0014.1 | GO:0004857 | enzyme inhibitor activity |
| Becon104Scf00044g1064.1 | GO:0046872 | metal ion binding |
| Becon104Scf07717g0004.1 | GO:0005576 | extracellular region |
| Becon104Scf03726g0023.1 | GO:0004866 | endopeptidase inhibitor activity |
| Becon104Scf03734g0005.1 | GO:0004190 | aspartic-type endopeptidase activity |
| Becon104Scf06110g0002.1 | GO:0000287 | magnesium ion binding |
| Becon104Scf01795g0013.1 | GO:0006952 | defense response |
| Becon104Scf04238g0017.1 | GO:0004869 | cysteine-type endopeptidase inhibitor activity |
| Becon104Scf06047g0004.1 | GO:0000287 | magnesium ion binding |
| Becon104Scf03734g0004.1 | GO:0004190 | aspartic-type endopeptidase activity |
| Becon104Scf10447g0003.1 | GO:0008152 | metabolic process |
| Becon104Scf01156g0006.1 | GO:0005515 | protein binding |
| Becon104Scf09326g0005.1 | GO:0005525 | GTP binding |
| Becon104Scf01496g0003.1 | GO:0000287 | magnesium ion binding |
| Becon104Scf00188g1001.1 | GO:0006952 | defense response |
| Becon104Scf03019g0002.1 | GO:0006629 | lipid metabolic process |
| Becon104Scf00224g1086.1 | GO:0003824 | catalytic activity |
| Becon104Scf08483g0005.1 | GO:0005515 | protein binding |
| Becon104Scf06072g0003.1 | GO:0000287 | magnesium ion binding |
| Becon104Scf14186g0005.1 | GO:0000287 | magnesium ion binding |
| Becon104Scf01013g0005.1 | GO:0005515 | protein binding |
| Becon104Scf05487g0001.1 | GO:0000786 | nucleosome |
| Becon104Scf00225g0001.1 | GO:0000786 | nucleosome |
| Becon104Scf00187g1006.1 | GO:0030145 | manganese ion binding |

Appendix 4. Annotations of genes significantly upregulated in *B. conchifolia* female flowers at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf08277g0003.1 | GO:0016021: | integral component of membrane |
| Becon104Scf02274g0021.1 | GO:0016021: | integral component of membrane |
| Becon104Scf06467g0003.1 | GO:0009765: | photosynthesis, light harvesting |
| Becon104Scf02248g0005.1 | GO:0015079: | potassium ion transmembrane transporter activity |
| Becon104Scf00791g0069.1 | GO:0003824: | catalytic activity |
| Becon104Scf16567g0002.1 | GO:0004185: | serine-type carboxypeptidase activity |
| Becon104Scf06944g0010.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf01401g0021.1 | GO:0008152: | metabolic process |
| Becon104Scf10618g0002.1 | GO:0005506: | iron ion binding |
| Becon104Scf00462g0046.1 | GO:0009765: | photosynthesis, light harvesting |
| Becon104Scf06558g0008.1 | GO:0008152: | metabolic process |
| Becon104Scf00411g0017.1 | GO:0005506: | iron ion binding |
| Becon104Scf01945g0005.1 | GO:0005506: | iron ion binding |
| Becon104Scf01250g0017.1 | GO:0003824: | catalytic activity |
| Becon104Scf15399g0013.1 | GO:0000272: | polysaccharide catabolic process |
| Becon104Scf03485g0014.1 | GO:0005515: | protein binding |
| Becon104Scf00011g3011.1 | GO:0016620: | oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor |
| Becon104Scf00740g0021.1 | GO:0008168: | methyltransferase activity |
| Becon104Scf03726g0001.1 | GO:0004866: | endopeptidase inhibitor activity |
| Becon104Scf00265g0006.1 | GO:0005576: | extracellular region |

Appendix 5. Annotations of genes significantly upregulated in *B. plebeja* female flowers at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf03726g0023.1 | GO:0004866: | endopeptidase inhibitor activity |
| Becon104Scf04238g0017.1 | GO:0004869: | cysteine-type endopeptidase inhibitor activity |
| Becon104Scf07717g0004.1 | GO:0005576: | extracellular region |
| Becon104Scf00044g1064.1 | GO:0046872: | metal ion binding |
| Becon104Scf05713g0001.1 | GO:0006952: | defense response |
| Becon104Scf00026g2043.1 | GO:0017148: | negative regulation of translation |
| Becon104Scf01804g0020.1 | GO:0006952: | defense response |
| Becon104Scf01795g0013.1 | GO:0006952: | defense response |
| Becon104Scf03074g0010.1 | GO:0004568: | chitinase activity |
| Becon104Scf00666g0023.1 | GO:0042742: | defense response to bacterium |
| Becon104Scf04016g0004.1 | GO:0005506: | iron ion binding |
| Becon104Scf02911g0018.1 | GO:0003824: | catalytic activity |
| Becon104Scf00034g0013.1 | GO:0016491: | oxidoreductase activity |
| Becon104Scf02545g0002.1 | GO:0004672: | protein kinase activity |
| Becon104Scf08483g0005.1 | GO:0005515: | protein binding |
| Becon104Scf01945g0005.1 | GO:0005506: | iron ion binding |
| Becon104Scf03361g0030.1 | GO:0004568: | chitinase activity |
| Becon104Scf00727g0062.1 | GO:0005506: | iron ion binding |
| Becon104Scf05040g0029.1 | GO:0008168: | methyltransferase activity |
| Becon104Scf00007g1014.1 | GO:0005506: | iron ion binding |
| Becon104Scf09326g0005.1 | GO:0005525: | GTP binding |
| Becon104Scf00065g0001.1 | GO:0006629: | lipid metabolic process |
| Becon104Scf05858g0002.1 | GO:0006952: | defense response |
| Becon104Scf01852g0030.1 | GO:0005515: | protein binding |
| Becon104Scf03712g0045.1 | GO:0008152: | metabolic process |
| Becon104Scf03712g0031.1 | GO:0008152: | metabolic process |
| Becon104Scf02000g0085.1 | GO:0004601: | peroxidase activity |
| Becon104Scf01156g0006.1 | GO:0005515: | protein binding |
| Becon104Scf00543g0081.1 | GO:0005507: | copper ion binding |
| Becon104Scf01322g0002.1 | GO:0003824: | catalytic activity |
| Becon104Scf03450g0001.1 | GO:0003824: | catalytic activity |
| Becon104Scf02349g0006.1 | GO:0003824: | catalytic activity |
| Becon104Scf00187g1001.1 | GO:0030145: | manganese ion binding |
| Becon104Scf09018g0001.1 | GO:0005507: | copper ion binding |
| Becon104Scf08445g0008.1 | GO:0004672: | protein kinase activity |
| Becon104Scf16139g0010.1 | GO:0004672: | protein kinase activity |
| Becon104Scf00253g1001.1 | GO:0005576: | extracellular region |

Appendix 6. Annotations of genes significantly upregulated in *B. conchifolia* leaf at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf06467g0003.1 | GO: 0009765 | photosynthesis, light harvesting |
| Becon104Scf00011g3011.1 | GO: 0016620 | oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor |
| Becon104Scf06144g0008.1 | GO: 0008270 | zinc ion binding |
| Becon104Scf00636g0019.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00740g0021.1 | GO: 0008168 | methyltransferase activity |
| Becon104Scf00843g0001.1 | GO: 0006869 | lipid transport |
| Becon104Scf04665g0002.1 | GO: 0005506 | iron ion binding |
| Becon104Scf02425g0005.1 | GO: 0004427 | inorganic diphosphatase activity |
| Becon104Scf16567g0002.1 | GO: 0004185 | serine-type carboxypeptidase activity |
| Becon104Scf01401g0021.1 | GO: 0008152 | metabolic process |
| Becon104Scf15399g0013.1 | GO: 0000272 | polysaccharide catabolic process |
| Becon104Scf06944g0008.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf06944g0010.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf07181g0006.1 | GO: 0003824 | catalytic activity |
| Becon104Scf00265g0006.1 | GO: 0005576 | extracellular region |
| Becon104Scf00343g0014.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf01945g0008.1 | GO: 0008152 | metabolic process |

Appendix 7. Annotations of genes significantly upregulated in *B. plebeja* leaf at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | Go term | Description |
|---|---|---|
| Becon104Scf06110g0002.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf06047g0004.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf03734g0005.1 | GO:0004190: | aspartic-type endopeptidase activity |
| Becon104Scf03734g0004.1 | GO:0004190: | aspartic-type endopeptidase activity |
| Becon104Scf01496g0003.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf05713g0001.1 | GO:0006952: | defense response |
| Becon104Scf00224g1086.1 | GO:0003824: | catalytic activity |
| Becon104Scf06072g0003.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf10447g0003.1 | GO:0008152: | metabolic process |
| Becon104Scf14186g0005.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf03019g0002.1 | GO:0006629: | lipid metabolic process |
| Becon104Scf00044g1064.1 | GO:0046872: | metal ion binding |
| Becon104Scf00843g0002.1 | GO:0006869: | lipid transport |
| Becon104Scf05956g0014.1 | GO:0004857: | enzyme inhibitor activity |
| Becon104Scf01795g0013.1 | GO:0006952: | defense response |
| Becon104Scf03074g0010.1 | GO:0004568: | chitinase activity |
| Becon104Scf02019g0001.1 | GO:0006629: | lipid metabolic process |
| Becon104Scf22098g0001.1 | GO:0009055: | electron carrier activity |
| Becon104Scf04238g0017.1 | GO:0004869: | cysteine-type endopeptidase inhibitor activity |
| Becon104Scf01156g0006.1 | GO:0005515: | protein binding |
| Becon104Scf08362g0005.1 | GO:0006508: | proteolysis |
| Becon104Scf09326g0005.1 | GO:0005525: | GTP binding |
| Becon104Scf00270g1050.1 | GO:0009055: | electron carrier activity |
| Becon104Scf02144g0001.1 | GO:0030145: | manganese ion binding |
| Becon104Scf02097g0016.1 | GO:0005779: | integral component of peroxisomal membrane |

Appendix 8. Annotations of genes significantly upregulated in *B. conchifolia* male flowers at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf01401g0021.1 | GO: 0008152 | metabolic process |
| Becon104Scf10618g0002.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00411g0017.1 | GO: 0005506 | iron ion binding |
| Becon104Scf01945g0005.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00187g1002.1 | GO: 0030145 | manganese ion binding |
| Becon104Scf07181g0006.1 | GO: 0003824 | catalytic activity |
| Becon104Scf06944g0010.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf06558g0008.1 | GO: 0008152 | metabolic process |
| Becon104Scf00559g0115.1 | GO: 0016787 | hydrolase activity |
| Becon104Scf01893g0002.1 | GO: 0003824 | catalytic activity |
| Becon104Scf00512g0018.1 | GO: 0008152 | metabolic process |
| Becon104Scf01155g0032.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf16567g0002.1 | GO: 0004185 | serine-type carboxypeptidase activity |
| Becon104Scf00265g0006.1 | GO: 0005576 | extracellular region |
| Becon104Scf00740g0021.1 | GO: 0008168 | methyltransferase activity |
| Becon104Scf03485g0014.1 | GO: 0005515 | protein binding |
| Becon104Scf01835g0003.1 | GO: 0030001 | metal ion transport |
| Becon104Scf01702g0015.1 | GO: 0008168 | methyltransferase activity |
| Becon104Scf00955g0062.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf00842g0029.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00011g3011.1 | GO: 0016620 | oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor |
| Becon104Scf04015g0024.1 | GO: 0005506 | iron ion binding |

Appendix 9. Annotations of genes significantly upregulated in *B. plebeja* male flowers at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf04238g0017.1 | GO:0004869: | cysteine-type endopeptidase inhibitor activity |
| Becon104Scf05713g0001.1 | GO:0006952: | defense response |
| Becon104Scf00044g1064.1 | GO:0046872: | metal ion binding |
| Becon104Scf03074g0010.1 | GO:0004568: | chitinase activity |
| Becon104Scf01795g0013.1 | GO:0006952: | defense response |
| Becon104Scf00265g1003.1 | GO:0005576: | extracellular region |
| Becon104Scf02000g0085.1 | GO:0004601: | peroxidase activity |
| Becon104Scf03726g0023.1 | GO:0004866: | endopeptidase inhibitor activity |
| Becon104Scf00065g0001.1 | GO:0006629: | lipid metabolic process |
| Becon104Scf04665g0002.1 | GO:0005506: | iron ion binding |
| Becon104Scf04204g0008.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf00098g0003.1 | GO:0008171: | O-methyltransferase activity |
| Becon104Scf01704g0025.1 | GO:0009507: | chloroplast |
| Becon104Scf24021g0003.1 | GO:0005576: | extracellular region |
| Becon104Scf09307g0014.1 | GO:0005506: | iron ion binding |
| Becon104Scf08483g0005.1 | GO:0005515: | protein binding |
| Becon104Scf00955g0062.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf04016g0004.1 | GO:0005506: | iron ion binding |
| Becon104Scf01156g0006.1 | GO:0005515: | protein binding |
| Becon104Scf09326g0005.1 | GO:0005525: | GTP binding |
| Becon104Scf05956g0014.1 | GO:0004857: | enzyme inhibitor activity |
| Becon104Scf03712g0045.1 | GO:0008152: | metabolic process |
| Becon104Scf02545g0002.1 | GO:0004672: | protein kinase activity |
| Becon104Scf00034g0013.1 | GO:0016491: | oxidoreductase activity |
| Becon104Scf03712g0031.1 | GO:0008152: | metabolic process |
| Becon104Scf00512g0018.1 | GO:0008152: | metabolic process |
| Becon104Scf00225g0001.1 | GO:0000786: | nucleosome |
| Becon104Scf08166g0006.1 | GO:0000287: | magnesium ion binding |

Appendix 10. Annotations of genes significantly upregulated in *B. conchifolia* petiole at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf06467g0003.1 | GO: 0009765 | photosynthesis, light harvesting |
| Becon104Scf00462g0046.1 | GO: 0009765 | photosynthesis, light harvesting |
| Becon104Scf00918g0088.1 | GO: 0005509 | calcium ion binding |
| Becon104Scf16567g0002.1 | GO: 0004185 | serine-type carboxypeptidase activity |
| Becon104Scf03478g0028.1 | GO: 0003824 | catalytic activity |
| Becon104Scf14875g0004.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf00028g0006.1 | GO: 0005515 | protein binding |
| Becon104Scf00011g3011.1 | GO: 0016620 | oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor |
| Becon104Scf07181g0006.1 | GO: 0003824 | catalytic activity |

Appendix 11. Annotations of genes significantly upregulated in *B. plebeja* petiole at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf00044g1064.1 | GO:0046872: | metal ion binding |
| Becon104Scf01510g0021.1 | GO:0008168: | methyltransferase activity |
| Becon104Scf04238g0017.1 | GO:0004869: | cysteine-type endopeptidase inhibitor activity |
| Becon104Scf05372g0011.1 | GO:0006952: | defense response |
| Becon104Scf01795g0013.1 | GO:0006952: | defense response |
| Becon104Scf09326g0005.1 | GO:0005525: | GTP binding |
| Becon104Scf03726g0023.1 | GO:0004866: | endopeptidase inhibitor activity |
| Becon104Scf01156g0006.1 | GO:0005515: | protein binding |
| Becon104Scf04015g0024.1 | GO:0005506: | iron ion binding |
| Becon104Scf00579g0076.1 | GO:0005506: | iron ion binding |
| Becon104Scf00225g0001.1 | GO:0000786: | nucleosome |
| Becon104Scf00955g0062.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf00293g0051.1 | GO:0008152: | metabolic process |
| Becon104Scf04204g0008.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf00224g1086.1 | GO:0003824: | catalytic activity |
| Becon104Scf08362g0005.1 | GO:0006508: | proteolysis |
| Becon104Scf00411g0046.1 | GO:0005506: | iron ion binding |

Appendix 12. Annotations of genes significantly upregulated in *B. conchifolia* root at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf00149g1094.1 | GO: 0004497 | monooxygenase activity |
| Becon104Scf14674g0001.1 | GO: 0004553 | hydrolase activity, hydrolyzing O-glycosyl compounds |
| Becon104Scf00531g0004.1 | GO: 0016491 | oxidoreductase activity |
| Becon104Scf00268g1024.1 | GO: 0004553 | hydrolase activity, hydrolyzing O-glycosyl compounds |
| Becon104Scf00071g2024.1 | GO: 0004497 | monooxygenase activity |
| Becon104Scf00265g0006.1 | GO: 0005576 | extracellular region |
| Becon104Scf01569g0071.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00189g0009.1 | GO: 0008270 | zinc ion binding |
| Becon104Scf01203g0075.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf03485g0014.1 | GO: 0005515 | protein binding |
| Becon104Scf00028g0006.1 | GO: 0005515 | protein binding |
| Becon104Scf00187g1002.1 | GO: 0030145 | manganese ion binding |
| Becon104Scf01556g0060.1 | GO: 0008152 | metabolic process |
| Becon104Scf16364g0011.1 | GO: 0005506 | iron ion binding |
| Becon104Scf01532g0004.1 | GO: 0008168 | methyltransferase activity |
| Becon104Scf00745g0057.1 | GO: 0005515 | protein binding |
| Becon104Scf01022g0028.1 | GO: 0004857 | enzyme inhibitor activity |
| Becon104Scf12899g0012.1 | GO: 0004601 | peroxidase activity |
| Becon104Scf06944g0010.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf00333g1049.1 | GO: 0045735 | nutrient reservoir activity |

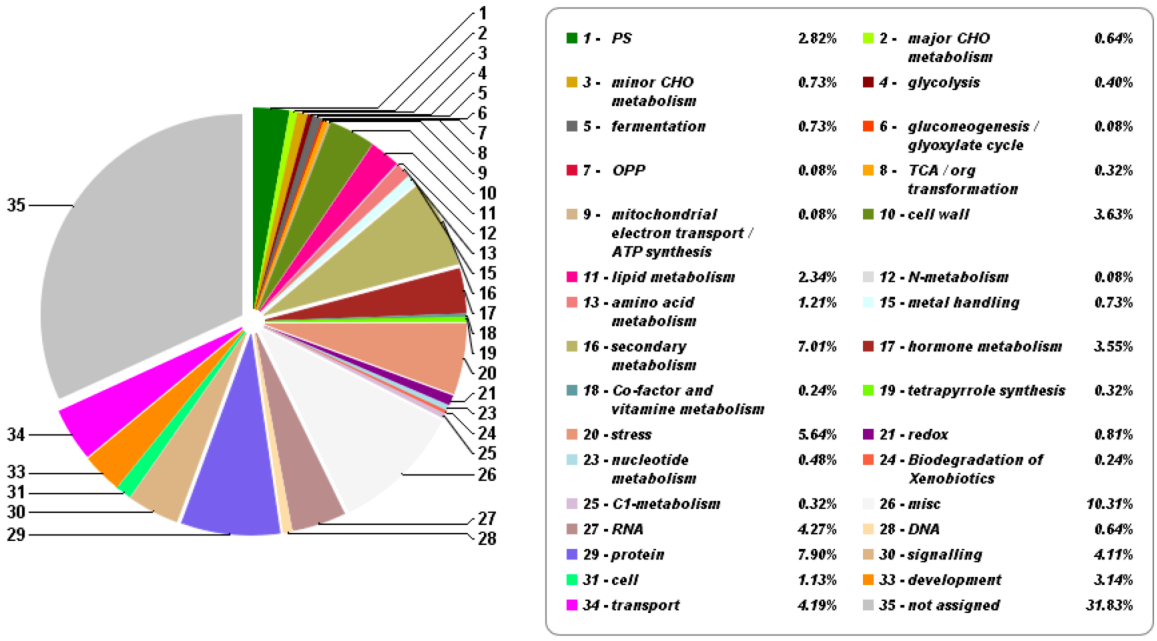
Appendix 13. Annotations of genes significantly upregulated in *B. plebeja* root at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf03726g0023.1 | GO:0004866: | endopeptidase inhibitor activity |
| Becon104Scf00044g1064.1 | GO:0046872: | metal ion binding |
| Becon104Scf01795g0013.1 | GO:0006952: | defense response |
| Becon104Scf09326g0005.1 | GO:0005525: | GTP binding |
| Becon104Scf05713g0001.1 | GO:0006952: | defense response |
| Becon104Scf03450g0001.1 | GO:0003824: | catalytic activity |
| Becon104Scf02349g0007.1 | GO:0003824: | catalytic activity |
| Becon104Scf00268g1024.1 | GO:0004553: | hydrolase activity, hydrolyzing O-glycosyl compounds |
| Becon104Scf01156g0006.1 | GO:0005515: | protein binding |
| Becon104Scf02349g0006.1 | GO:0003824: | catalytic activity |
| Becon104Scf00225g0001.1 | GO:0000786: | nucleosome |
| Becon104Scf02095g0022.1 | GO:0005515: | protein binding |
| Becon104Scf00553g0047.1 | GO:0005509: | calcium ion binding |
| Becon104Scf03712g0045.1 | GO:0008152: | metabolic process |
| Becon104Scf05735g0002.1 | GO:0080019: | fatty-acyl-CoA reductase |
| Becon104Scf06663g0019.1 | GO:0000287: | magnesium ion binding |
| Becon104Scf00483g0010.1 | GO:0004672: | protein kinase activity |
| Becon104Scf00224g1086.1 | GO:0003824: | catalytic activity |
| Becon104Scf03321g0014.1 | GO:0003677: | DNA binding |
| Becon104Scf00034g0013.1 | GO:0016491: | oxidoreductase activity |
| Becon104Scf00111g1017.1 | GO:0005618: | cell wall |
| Becon104Scf04194g0025.1 | GO:0005507: | copper ion binding |
| Becon104Scf05956g0014.1 | GO:0004857: | enzyme inhibitor activity |
| Becon104Scf01577g0014.1 | GO:0005509: | calcium ion binding |
| Becon104Scf00829g0030.1 | GO:0005509: | calcium ion binding |
| Becon104Scf16139g0010.1 | GO:0004672: | protein kinase activity |

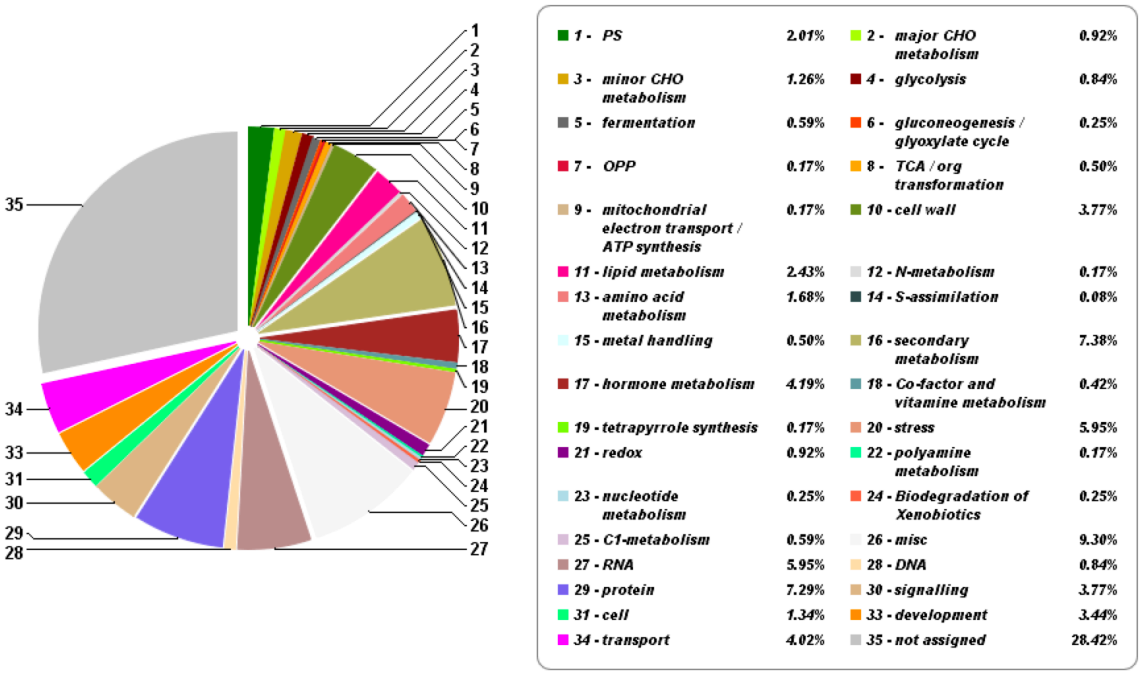
Appendix 14. Annotations of genes significantly upregulated in *B. conchifolia* vegetative bud at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.

| Locus name | GO term | Description |
|---|---|---|
| Becon104Scf00817g0031.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf00843g0001.1 | GO: 0006869 | lipid transport |
| Becon104Scf06467g0003.1 | GO: 0009765 | photosynthesis, light harvesting |
| Becon104Scf16567g0002.1 | GO: 0004185 | serine-type carboxypeptidase activity |
| Becon104Scf00265g0006.1 | GO: 0005576 | extracellular region |
| Becon104Scf03485g0014.1 | GO: 0005515 | protein binding |
| Becon104Scf03281g0010.1 | GO: 0005515 | protein binding |
| Becon104Scf06944g0010.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf06944g0008.1 | GO: 0000287 | magnesium ion binding |
| Becon104Scf07181g0006.1 | GO: 0003824 | catalytic activity |
| Becon104Scf00187g1002.1 | GO: 0030145 | manganese ion binding |
| Becon104Scf00579g0022.1 | GO: 0005506 | iron ion binding |
| Becon104Scf00003g0020.1 | GO: 0003824 | catalytic activity |
| Becon104Scf00187g1006.1 | GO: 0030145 | manganese ion binding |
| Becon104Scf00343g0014.1 | GO: 0006629 | lipid metabolic process |
| Becon104Scf10637g0009.1 | GO: 0006952 | defense response |

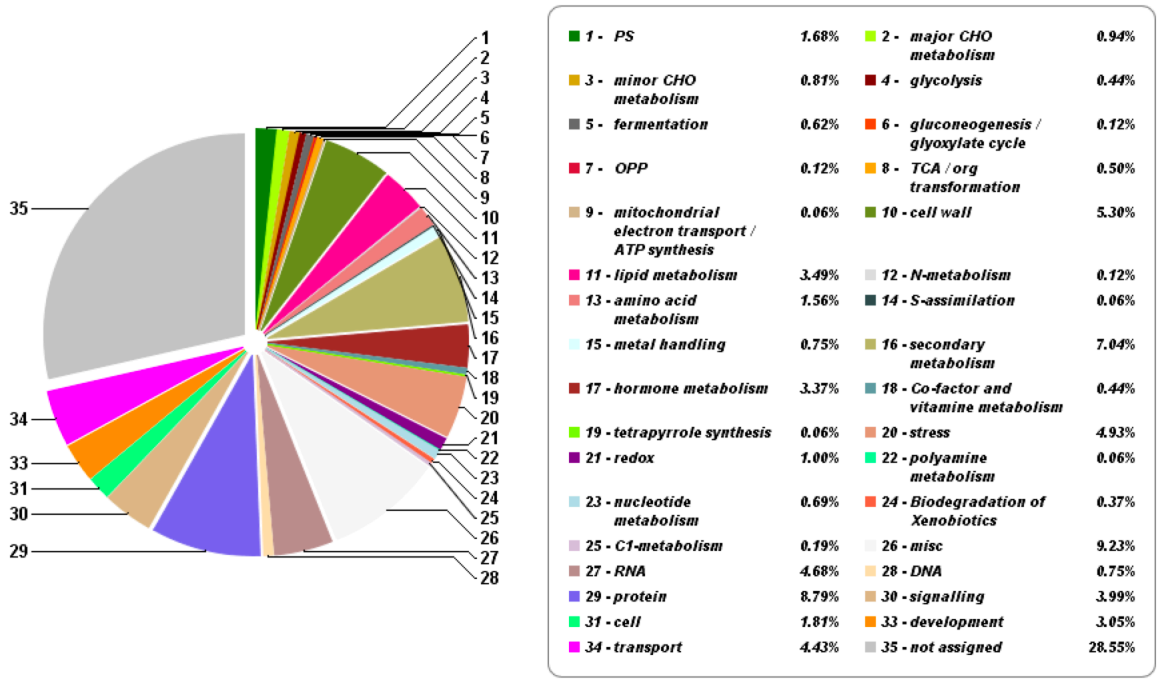
Appendix 15. Annotations of genes significantly upregulated in *B. plebeja* vegetative bud at alpha = 0.05 with FDR adjusted p-values. Annotations are transferred from the *B. conchifolia* draft genome functional annotations.



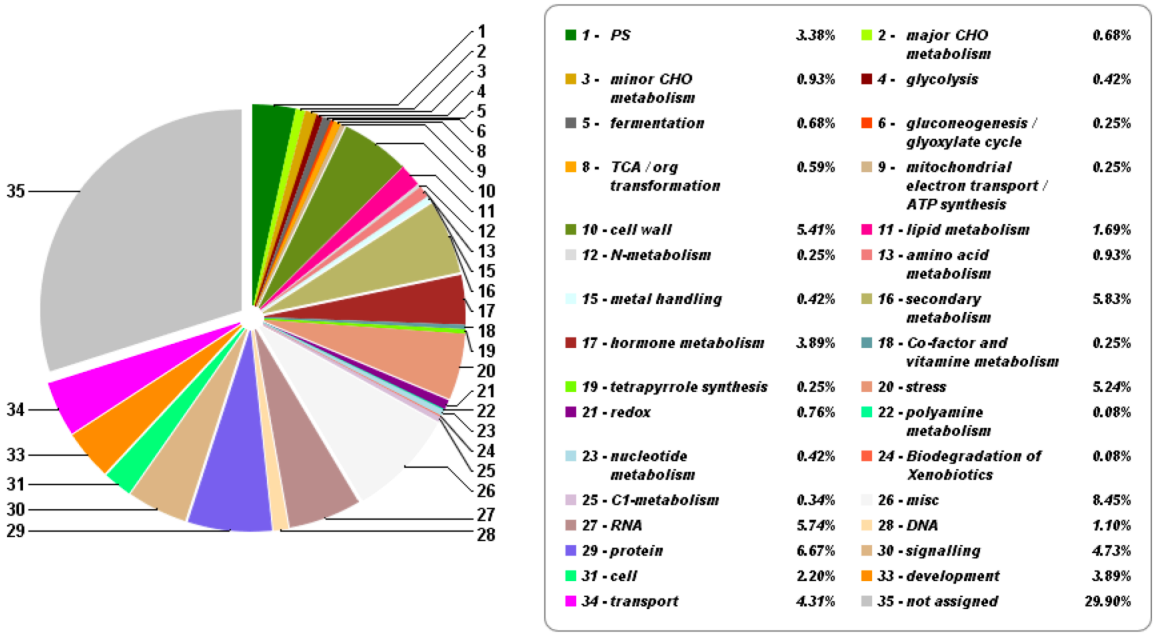

Appendix 16. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* female flower



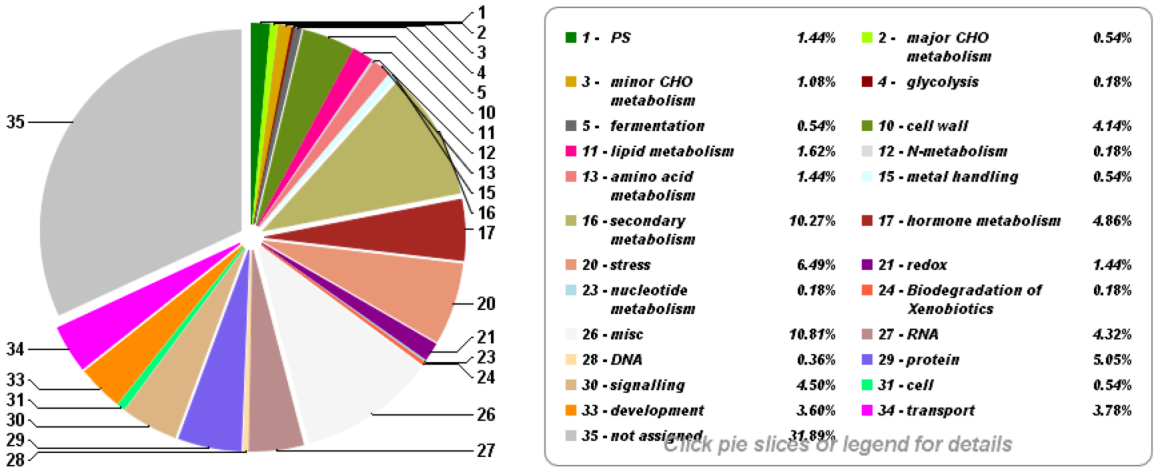

Appendix 17. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* leaf



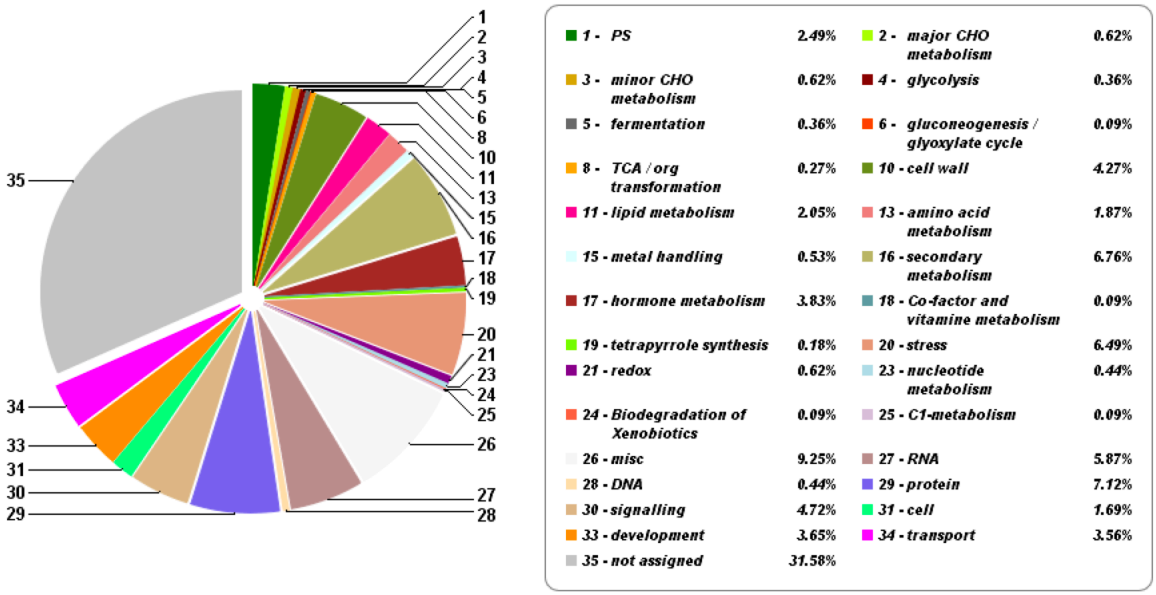

Appendix 18. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* male flower




Appendix 19. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* petiole




Appendix 20. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* root




Appendix 21. Mercator pie chart showing distribution of genes differentially expressed between *B. conchifolia* and *B. plebeja* vegetative bud