Genomic analyses of gene flow and selection during diversification in *Senecio*

Owen Gregory Osborne

The Queen's College, Oxford

A thesis submitted for the degree of Doctor of Philosophy, Trinity term,

2015. Under the supervision of Dr. Dmitry A Filatov and Dr. Steve Kelly

Abstract

Genomic analyses of gene flow and selection during diversification in *Senecio*

Owen Gregory Osborne, The Queen's College. A thesis submitted for the degree of Doctor of Philosophy, Trinity term, 2015.

Diversifying selection and gene flow were traditionally viewed as antagonistic forces, with diversifying selection promoting adaptation and speciation, and gene flow opposing them. However, their roles are far more complex than this. While gene flow can prevent speciation or initiate despeciation, it can also generate new hybrid species. Similarly, while adaptive divergence can be wiped out by gene flow, new adaptive variation can be introduced via hybridisation. The relative frequency of these outcomes, and indeed the frequency of gene flow and diversifying selection in general are largely unknown. This thesis illuminates these questions through evolutionary genomic analyses focussed on a recently diverged group of ragworts (*Senecio*). The Mediterranean *Senecio* species-complex contains several cases of hybrid speciation, as well as two species, *S. aethnensis* and *S. chrysanthemifolius*, which are a potential case of ecological speciation with gene flow, having adapted to high and low altitude habitats respectively on Mount Etna. However, their demography was previously un-studied. I first show that *S. aethnensis* and *S. chrysanthemifolius* diverged recently, at a time that coincides with the growth of Mount Etna to the altitudes which separate the species today, have experienced significant gene flow following their split, and are likely to be sister species, bolstering the hypothesis of ecological speciation. I further demonstrate that gene flow is common in the wider clade, and pinpoint multiple cases of gene flow amongst them. Finally, I identify several genes under positive selection in the clade, and show that the proportion of genes under selection is high relative to many other plant genera. The results further establish *Senecio* species as an invaluable model system for the study of diversification with gene flow, and suggest that high levels of gene flow and selection are features of their evolution, a situation which may prove to be common in plants.

Acknowledgements

The years I've spent completing my DPhil have included the most interesting, happiest, and the most terrifying moments of my life, so there are a great number of people I have to thank for steering my time in Oxford towards the former two.

I would first like to thank my supervisors, Dmitry Filatov and Steve Kelly, for providing guidance when necessary but allowing me to work independently and act on my own ideas. I would also like to thank the Gatsby Charitable Foundation for funding my DPhil, and to all the great people I've met there, for making the experience so much richer.

The Filatov lab has always been an interesting place to work, and I'd like to thank everyone I've had the pleasure of sharing an office with. Particular thanks goes to Mark Chapman, Mike Chester, Bruno Nevado and Alex Papadopulos without whom I would certainly be a poorer scientist (and may not have even finished). For introducing me to the Filatov lab in the first place, and for thereafter being enthusiastic collaborators, I'd like to thank Simon Hiscock and Tom Batstone. I spent a fantastic two months in Sheffield in my final year working in Patrik Nosil's lab, so I'd like to thank Patrik and everyone else in the Nosil lab. I had a great time and learnt a lot. And I'd love to return one day.

In addition to colleagues, there are a number of friends and family members who have made my DPhil both possible and enjoyable. Science is my second career, and I would probably still be a builder if it wasn't for two people in particular, so I'd like to thank Ross Cuntlipps for putting the idea of returning to education into my head, and Debbie Cohen for making me do it. I've made so many good friends in Oxford that it would be impossible to thank them all, but some deserve special mention. Y-ling Chi has been the most positive and consistently good influence on me during the last four years. Weighing against that good influence, Hector Page, Dan Yin and Suvi Honkanen have probably been my most consistent partners in crime, although there are many, many more. A special thanks also has to go to Frau Doctor Carolin Schultz and Cal "the frau plough" Musto, for giving me a sofa to sleep on and mothering me through my final hellish week of homeless thesis writing. I'd also like to thank my mum and step-dad: whether I was signing on, or completing a DPhil at Oxford, they have always been incredibly supportive. Finally, I want to thank my better half Sarah, for always knowing how to pull me out of the deepest states of DPhil fatigue and put a smile on my face.

Table of contents

Abstract ***1***

Acknowledgements ***2***

Chapter 1: General introduction ***7***

1.1 What is a speciation? What is a species? 7

1.2 What are the roles of gene flow and selection during adaptation and speciation? 9

1.3 Phylogenetics in the light of gene flow 12

1.4 The pattern and prevalence of selection during species diversification 13

1.4 Study systems 15

Chapter 2: Demographic inference and genome-wide analysis of selection pressures in Senecio. ***19***

Preface **19**

2.1 Introduction **20**

2.2 Materials and Methods **22**

2.2.1 Sequencing 22

2.2.2 Dataset preparation 23

2.2.3 Transcriptome annotation and validation 25

2.2.4 Testing the mode of speciation 26

2.2.5 Divergence analysis 27

2.2.6 Factors affecting the genome-wide selective landscape. 28

2.3 Results **30**

2.3.1 Transcriptome sequencing, validation and characterisation 30

2.3.2 Testing the mode of speciation 32

2.3.3 Transcriptome-wide analyses of d*N*/d*S* 36

2.4 Discussion **39**

2.4.1 Characterisation of the *Senecio* transcriptomes. 39

2.4.2 Recent speciation may have been driven by the growth of Mount Etna 40

2.4.3 The genome wide landscape of selection in *Senecio* 44

2.4.4 Conclusions 47

Chapter 3: Maintenance of species boundaries despite ongoing gene flow in Mediterranean Senecio. ***48***

Preface **48**

3.1 Introduction **49**

3.2 Materials and Methods **52**

3.2.1 Seed collection, plant growth and sequencing 52

3.2.2 Dataset preparation 53

3.2.3 Phylogenetic inference 55

3.2.4 Tests for introgression 57

3.3 Results **58**

3.3.1 Dataset 58

3.3.2 Phylogenetic inference 58

3.3.3 Detection of introgression 60

3.4 Discussion **62**

3.4.1 Species level phylogeny and gene tree-species tree incongruence 62

3.4.2 Introgression is widespread in the group 65

3.4.3 Conclusions 68

Chapter 4: The rate of positive selection in Senecio relative to other plant taxa. ***72***

Preface **72**

4.1 Introduction **73**

4.2 Materials and methods **75**

4.2.1 Data collection 75

4.2.2 Dataset preparation 79

4.2.3 Phylogenetic analysis 81

4.2.4 Detection of positive selection 82

4.2.5 Functional enrichment analysis 84

4.3 Results **84**

4.3.1 Sequence data processing 84

4.3.2 Phylogenetic analysis 87

4.3.3 Dataset filtering 91

4.3.4 Genes under positive selection in *Senecio* 92

4.3.5 Rates of selection in *Senecio* relative to other plant genera 93

4.4 Discussion **97**

4.4.1 Validity of results 97

4.4.2 A diverse collection of protein-coding genes are under selection in *Senecio* 98

4.4.3 Selective regimes in orthologues are highly correlated in different plant taxa 99

4.4.4 *Senecio* has a high rate of positive selection relative to many plant genera 100

4.4.5 Do plants experience high levels of positive selection in general? 103

4.4.6 Conclusions 104

Chapter 5: General discussion ***104***

5.1 *Senecio aethnensis* and *Senecio chrysanthemifolius* - an example of ecological speciation with gene flow? 104

5.2 The Mediterranean *Senecio* species-complex as a model for diversification with gene flow 109

5.3 The genomic basis for adaptation in *Senecio* 112

5.4 Conclusions 114

References ***117***

Chapter 1: General introduction

Every taxon, however large and diverse, derived at some point in the past from a single, homogenous population. This path from panmixis to reproductive isolation, and from uniformity to difference, is the central processes which evolutionary biology attempts to explain. The overall process must be understood at multiple scales. The emergence via selection or drift, and eradication via gene flow or extinction, of differences between interconnected populations; the speciation process itself, in which a discontinuity evolves within a previously continuously varying species and reproductive isolation becomes established; and the adaptation and diversification of taxa over time to fill available niche-space. There are many gaps in knowledge at all these scales of the process of diversification. This chapter will outline what is known and unknown about the processes of speciation and adaptation, with particular regard to the prevalence and roles of gene flow and selection during these processes; and will detail the general aims of the thesis in elucidating these questions through genomic studies of plant evolution, focussed on the genus *Senecio*.

1.1 What is a speciation? What is a species?

Speciation is the division of a single species into two, so to clearly define this process it is essential to briefly state what is meant by a species at all. The endeavour to impose a classification system on the inherent complexity of biology, and the specific question of what constitutes a species, has resulted in one of the major discourses in 20th century evolutionary biology (Hey 2006). The oldest definitions of species were, like those of higher taxonomic classifications, relatively arbitrary, and were effectively a “kind” of organism.

Mayr's (1942) 'Biological Species Concept' (BSC) was revolutionary in that it rendered species a real biological phenomenon, and also suggested evolutionary pathways which might lead to their genesis.

The BSC defines species as "groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups" (Mayr 1957). Thus, the central criterion of the BSC is complete reproductive isolation (RI), and under strict interpretations, any breach of a species' integrity by gene flow with a variant form, violates RI and relegates both forms to sub-specific status. This definition is neat and clearly identifies a real biological phenomenon, that of the cessation of gene flow between previously interbreeding groups, but it fails to capture the complex reality of certain groups of organisms. "Species" in many taxa, especially plants such as oaks (*Quercus* spp; Curtu et al., 2007) and ragworts (*Senecio* spp*;* Comes & Abbott, 2001), are known to exchange genes frequently but retain highly distinct 'species'-specific differences. The criterion of RI demotes many groups of species, which are clearly morphologically distinct and discontinuous, to the status of being one single species because of occasional episodes of - or even the unrealised potential for - gene flow. Conversely, if the BSC is blindly applied, a polyploidisation event which causes reproductive isolation from the diploid parents instantly creates a new species, despite the possibility of no ecological or morphological differences from its parents. Similar complications arise when considering cases of hybrid speciation. Even if the progenitor species are reproductively isolated in a conventional sense then some forms of hybrid speciation may be possible. For example diploid *Senecio squalidus* and tetraploid *S. vulgaris* are reproductively isolated because they produce sterile triploid progeny, known as *Senecio x baxteri*. Occasionally, however, *Senecio x*

baxteri individuals can undergo genome doubling which results in a hexaploid fertile and self-compatible individual which is reproductively isolated from both parent species (Lowe & Abbott 1996; Hegarty et al. 2006). Further complications abound. For example multiple such allopolyploids may not be reproductively isolated from each other but may be isolated from their diploid progenitors, so this raises the problem of whether they should be classed as a single species despite being polyphyletic in origin (Lowe & Abbott 2004; Wyatt et al. 1988). Moreover, in asexual species, RI is clearly irrelevant, yet many asexual taxa still form discontinuous phenotypic and genotypic clusters just as sexual taxa do (Birky & Barraclough 2009).

The crystallisation of the importance of RI in the speciation process was clearly a crucial step, but focussing on RI as a Boolean property which is essential for species to be regarded as separate entities clearly oversimplifies the variety of modes of speciation in nature. In this thesis, speciation is regarded as an ongoing process rather than a decisive event, and species are regarded as being somewhere along a continuum of speciation states, from being completely panmictic and homogenous, to completely reproductively isolated with a clear discontinuity between them, without seeking to define when they have crossed the "finishing line" into full specific status.

1.2 What are the roles of gene flow and selection during adaptation and speciation?

Gene flow and selection may be the most important forces guiding speciation and adaptation. A simplistic view of their influence is that divergent selection promotes adaptation and speciation, and that gene flow retards local adaptation and must cease before speciation can occur (Mayr 1963; Wright 1931). This situation can indeed be the

case, and there are many examples of gene flow hampering adaptation and speciation (Bridle & Vines 2007). Indeed, in some cases, reinforcement of RI in areas of sympatry evolves for this reason (Silvertown et al. 2005; Hopkins & Rausher 2011). This is not to say that gene flow between diverging species is necessarily maladaptive, however. There are multiple examples of positive selection on introgressed genes (Huerta-Sánchez et al. 2014; Sun et al. 2012). It has even been suggested that increased rates of hybridisation following the colonisation of new habitats may predispose newly colonising species to adaptive divergence (Seehausen, 2004). This "hybrid swarm theory" predicts that as species diversification proceeds, hybridisation between its recently diverged species produces more variation on which selection can act and accelerates their diversification.

The notion that gene flow must be externally halted before speciation can be achieved (the allopatric speciation model; Mayr 1963) has also been largely superseded. Linked to the realisation of the importance of RI in speciation was the emphasis of geography, and the division of speciation into allopatric, sympatric and parapatric geographic modes (Coyne 1994). Allopatric speciation was seen as being overwhelmingly predominant, or even the only plausible mode of speciation, because geographic separation externally prevented gene flow (Mayr 1963; Coyne & Orr 2004; Bolnick & Fitzpatrick 2007). Speciation in the presence of gene flow was considered unlikely because, based on contemporary knowledge, the combined forces of gene flow and recombination would break down linkage disequilibrium between differentially adapted genes too effectively for divergence to build up (Bolnick & Fitzpatrick 2007). However, several compelling cases of sympatric speciation (Schliewen et al. 1994; Filchak et al. 2000; Savolainen et al. 2006) coupled with theoretical advances (Kondrashov & Kondrashov 1999; Wu 2001; Doebeli & Dieckmann

2002) eventually made this position untenable. Divergent processes driven by natural selection in the presence of gene flow, rather than processes arising as a by-product of the cessation of gene flow, are sufficient to cause the origin of new lineages (Pinho & Hey 2010). It is now accepted that speciation with gene flow occurs (Rundle & Nosil 2005; Pinho & Hey 2010; Smadja & Butlin 2011) and the discussion has moved to its relative frequency compared to other modes of speciation and what genetic mechanisms can explain it (Feder & Nosil 2010).

The interplay between gene flow and selection during speciation with gene flow has profound consequences for the structure of genomic divergence. In his influential review, Wu (2001) pointed out that, while the gene is usually considered the unit of adaptation, concepts of species, speciation and gene flow have often been viewed from an organismal, and therefore whole-genome, perspective. However, differential adaptation to dissimilar environments, which is likely to be involved in ecological speciation with gene flow (Smadja & Butlin 2011), may not be expected to involve all, or even a large proportion of, the genes in the genome. Wu (2001) described a model whereby genomes are porous during speciation, with divergently selected loci forming 'islands of speciation' whereas other loci are free to pass between populations. This process leads to a 'mosaic genome' with respect to divergence in incipient species, and regions of the genome containing loci under strong divergent selection and lower rates of gene flow can be identified by their elevated genetic divergence. Early empirical evidence for a 'mosaic' or 'porous' genome model came from a study in mosquitoes in which two divergent *Anopheles gambiae* forms were genetically differentiated (i.e. showed reduced gene flow) in only three regions encompassing 2.8 Mb (Turner et al., 2005). Since then, numerous studies have found highly variable levels of

genomic differentiation in incipient species (Renaut et al. 2012; Jones et al. 2012; Poelstra et al. 2014) and regions of high differentiation are often taken to represent genes under divergent selection, which are consequently shielded from gene flow (although selection following speciation can also create similar genomic signatures so careful interpretation is needed; Cruickshank & Hahn 2014). Detailed knowledge of the rate of gene flow following the initiation of divergence, and variation in selective regimes between genomic regions in diverging taxa are needed to both confirm divergence with gene flow, and elucidate the genomic basis for the process. Therefore recent or ongoing examples of speciation are needed. Chapter 2 of this thesis investigates the speciation process in such an example, *Senecio aethnensis* and *S. chrysanthemifolius*, two species which are closely related, are known to hybridise promiscuously, but show dramatic phenotypic divergence, and are thus a possible example of recent but incomplete ecological speciation with gene flow.

1.3 Phylogenetics in the light of gene flow

The realisation that gene flow can occur during species diversification has profound consequences for phylogenetics. The traditional representation of a clade's evolutionary history is a branching phylogenetic tree. However, if some species are the result of hybrid speciation or if introgression leads to differing phylogenetic histories across the genome then this is patently inaccurate. Introgression, the movement of genes from the gene pool of one species to another, through repeated hybridisation and backcrossing (Anderson 1949), may be extremely prevalent in nature (and particularly in plants; Whitney et al. 2010). This is often the case despite divergent local adaptation and strong pre- and post-zygotic reproductive barriers (Lawton-Rauh et al. 2007; Chapman & Abbott 2010; Sambatti et al. 2012; Cui et al. 2013). Such introgressive hybridisation, as well as incomplete lineage

sorting, can cause significant phylogenetic incongruence between different genomic regions in a species complex and this can lead to difficulty in estimating the species phylogeny. Rather than being viewed simply as an inconvenience for phylogenetic inference, however, the various incongruent phylogenetic histories observed between loci should correctly be seen to represent a more accurate account of the history of a clade than any single bifurcating phylogeny when hybridisation has occurred and where ancestral polymorphism has not reached fixation amongst descendant lineages (Rokas & Carroll 2006). Chapter 3 of this thesis examines these issues in the Mediterranean *Senecio* species complex, by determining the "average" phylogenetic history of the species, identifying the frequency of introgression between them, and determining to what extent this introgression may have created phylogenetic incongruence between different genomic regions.

1.4 The pattern and prevalence of selection during species diversification

What of the force assumed to typically oppose gene flow, diversifying selection? Diversifying selection and consequent adaptation is generally assumed to be the main driver of biological diversity, at least in terms of major phenotypic differences between species. However, at the molecular level the role, pattern and prevalence of selection are far less clear. Population geneticists commonly divide selection into purifying selection (selection to eliminate deleterious mutations) and positive selection (selection to increase the frequency of beneficial mutations). Positive selection may drive a beneficial mutation to fixation (directional selection) or favour a polymorphism (balancing selection). Kimura's neutral theory of molecular evolution challenged the previous assumption that most genetic divergence is adaptive (Kimura 1983). The theory suggests that most variation is

selectively neutral, most mutations are deleterious and are quickly purged by purifying selection, and that positive selection is rare at the molecular level. The more recent nearly neutral theory of molecular evolution views most variation as weakly deleterious or weakly adaptive, but with a strength of selection below the threshold required to be effectively purged or fixed (Ohta 1992).

Nevertheless, population genetic evidence of positive selection has been found at the molecular level and many statistical approaches have been developed to detect it, as well as to measure the strength of both positive and purifying selection (Hudson et al. 1987; McDonald & Kreitman 1991; Nielsen et al. 2005; Nielsen & Yang 1998; Tajima 1989). Often these approaches have been applied to candidate genes suspected of being under strong positive selection, such as those involved in host-pathogen arms races (Sironi et al. 2015). However, functional knowledge of genes is often poor, and as clades diversify, the many different selection pressures which affect them may be unknown. Thus potential targets of positive selection are rarely known *a priori*. The massive recent improvements in sequencing throughput allows an opposite approach, in which large proportions of the genome can be "scanned" for evidence of selection with no initial hypothesis regarding the loci under selection. This tactic, referred to as "reverse ecology" by Li et al. (2008), promises insight into multiple fundamental questions concerning the prevalence of positive selection during speciation and diversification.

In addition to identifying the loci under positive selection, which may suggest hypotheses for the selective forces driving speciation and adaptation, large unbiased sequence datasets can test evolutionary hypotheses regarding the role of selection on a genome-wide scale. These could include determining the proportion of the genome evolving

neutrally, under purifying or under positive selection (Boyko et al. 2008; Enard et al. 2014); asking to what extent the same genes are under similar selective regimes in different species (Kane et al. 2011; Soria-Carrasco et al. 2014); or diverse theories on the predictability of evolution, such as whether coding or regulatory regions are more often involved in adaptation (Jones et al. 2012). Chapter 4 of this thesis uses a reverse ecology approach to identify protein-coding genes under selection in *Senecio*, compares the proportion of genes under positive selection in *Senecio* to those in multiple other plant genera, and investigates the extent of parallel selective processes acting on orthologues between *Senecio* and these other taxa.

1.4 Study systems

This thesis focusses on the genus *Senecio*, but also uses data from other plant genera for comparison. Firstly, a single instance of possible ecological speciation with gene flow is characterised in detail using *Senecio aethnensis* and *S. chrysanthemifolius*, two species which may be in the early stages of speciation in response to altitude (Chapter 2). Secondly, the dynamics of gene flow during clade diversification are investigated by broadening the study system to include several more species in the Mediterranean clade which includes *S. aethnensis* and *S. chrysanthemifolius* (Chapter 3). Finally to attempt to determine the prevalence and patterns of natural selection during clade diversification, the genus *Senecio* as well as several other broadly phylogenetically distributed plant genera form the focus.

Senecio aethnensis and *S chrysanthemifolius* have several characteristics which suggest they may be an example of recent, ecological speciation. *Senecio aethnensis* Jan ex DC. grows at high altitudes on Mount Etna (>1600 m), whereas *S. chrysanthemifolius* Poir. is largely confined to the volcano's lower slopes (<1000m; as well as other regions of Sicily

and a small population on the Southernmost tip of mainland Italy; Chapman et al. 2005). At intermediate altitudes, where the ranges of these two species approach one another, there is a stable hybrid zone (Chapman et al. 2005; James & Abbott 2005). While 'purer' forms of *S. aethnensis* and *S. chrysanthemifolius* differ in several phenotypic characteristics, such as leaf shape, capitulum (inflorescence) size, and ray flower size, and are regarded by botanists as 'good species', plants in the hybrid zone show a range of intermediate phenotypes which track an altitudinal cline (James & Abbott 2005; Brennan et al. 2009). Ecological factors that are likely to differ between their high and low altitude habitats include: UV-light exposure, temperature and water availability (James & Abbott 2005; Brennan et al. 2009) and germination temperature in the greenhouse has been shown to correlate with the altitude of their source population, likely as a result of selection (Ross et al. 2012). Given the high inter-fertility of *S. aethnensis* and *S. chrysanthemifolius* (Chapman et al, 2005), this system provides an excellent platform to determine the rate of gene flow following divergence, examine the selective forces which maintain these species' considerable phenotypic divergence in the face of what may be substantial gene flow, and to identify the genomic regions responsible.

Clinal analysis of the *Senecio aethnensis – S. chrysanthemifolius* hybrid zone using quantitative traits and a small (n=13) cohort of molecular markers, showed that the hybrid zone is shaped by gene flow and selection against hybrids (Brennan et al., 2009). Differences between the clines for certain quantitative traits suggested that environmental selection is largely responsible for structuring trait differentiation across the hybrid zone. In contrast, at finer spatial scales gene flow and intrinsic selection against hybrids may be more important than environmental selection. Prior to the work in this thesis however, all

studies of genetic divergence in the system have been limited by the small number of loci which they have used (Brennan et al. 2009; Muir et al. 2013). There is also evidence of a modest amount of both pre- and post-zygotic reproductive isolation between the species. The species have different flowering times, both in nature and in the greenhouse (Ross 2010) and a significant proportion of second generation hybrids are inviable (Brennan et al. 2014; Chapman et al. 2016).

For the analyses in Chapters 3 and 4, eight diploid species of *Senecio* were used. This included six from the Mediterranean species complex (*S. aethnensis* Jan ex DC., *S. chrysanthemifolius* Poir., *S. leucanthemifolius* Poir., *S. gallicus* Vill., *S. glaucus* L. and *S. vernalis* Waldst. & Kit.) and two outgroups (*S. madagascariensis* Poir. and *S. flavus* (Decne.) Sch. Bip.). The Mediterranean *Senecio* species complex provides a classic illustration of the diverse modes by which diversification can progress (Comes & Abbott 2001). It contains examples of a stable hybrid zone (Brennan et al., 2009), and both allopolyploid (Kadereit et al. 2006; Lowe & Abbott 1996, 2004; Pelser et al. 2012) and homoploid hybrid speciation (James & Abbott 2005). Therefore many of the species are known to be capable of hybridising and in addition to examples of hybrid speciation, some evidence for the more subtle process of introgressive hybridisation has been found (Comes & Abbott 1999, 2001; Chapman & Abbott 2010), although this has received less attention (but see Coleman & Abbott 2003; Chapman & Abbott 2005, which discounted hypotheses of introgression in the group). Despite the possibility of hybridisation between many of the species, they are highly phenotypically distinct and occur in a wide range of environments, including desert, alpine, steppe, rocky volcanic and coastal Mediterranean habitats, so are likely to be

divergently ecologically adapted. Thus, the system represents an opportunity to examine how clade divergence proceeds in the presence of both gene flow and selection.

In addition to the *Senecio* species studied in Chapter 4, eight other Angiosperm genera were examined. The genera used were *Flaveria, Glycine, Helianthus, Linum, Oryza, Populus, Silene* and *Solanum*. The genera used were determined by the published data available. Nevertheless, the resulting set of genera represents a taxonomically diverse selection of Angiosperms including monocots, and eudicots and spanning seven plant families. Thus, the selected genera provide a superb dataset for the study of the rate of selection during diversification in general and how this varies and is similar between taxa.

Overall, the thesis aims to shed light on several interconnected questions in the biology of speciation and divergence and the prevalence of gene flow and selection during these processes. This is achieved through the analysis of large high-throughput sequence datasets from plants representing a view of these processes from several scales: from the early stages of speciation, to a larger clade which has significantly diversified but may still be capable of hybridisation, and finally a group of multiple taxa representing a broad spectrum of scenarios of clade diversification.

Chapter 2: Demographic inference and genome-wide analysis of selection pressures in *Senecio*.

Preface

This chapter was published in a slightly different form as “Rapid speciation with gene flow following the formation of Mount Etna” in the journal *Genome Biology and Evolution* (volume 5, issue 9, pp. 1704-1715). All experimental work and writing is my own with the following exceptions. Dmitry Filatov, Thomas Batstone and Simon Hiscock and two anonymous reviewers provided comments and corrections on the manuscript. Mark Chapman grew the plants used and extracted RNA for sequencing, Michael McKain ran the paralogue identification analysis.

2.1 Introduction

Changes in the environment which a species occupies can create new niches which drive ecological species divergence without the immediate cessation of gene flow (Rundle & Nosil 2005). While the speed of this process in general is unknown, there are multiple lines of evidence for at least the early stages of speciation, before complete reproductive isolation has been reached, occurring very quickly in many cases (Bearhop et al. 2005; Feder et al. 1994; Silvertown et al. 2005). How common this process is, how quickly divergence can exploit new ecological opportunities, and how much gene flow can be tolerated during divergence, remain open questions, however (Hendry 2007; Nosil et al. 2009; Pinho & Hey 2010).

Recent technological advances are revolutionising the study of speciation genomics by allowing non-model organisms to be studied on a genome-wide scale (Rokas & Abbot 2009; Ekblom & Galindo 2011; e.g Jones et al. 2012; The Heliconius Genome Consortium et al. 2012; Poelstra et al. 2014). It has become feasible to choose study species on the basis of their ecological suitability for answering specific biological questions, rather than, as before, the existence of genomic resources. Furthermore, a wider range of study systems makes the discrimination of system-specific phenomena, from those which are more general, much less challenging, so the need for a greater number of evolutionary study systems with sufficient genomic resources is significant. Specific to the field of speciation research, examples of 'speciation in action' can now be chosen, in which the speciation process took place very recently or is on-going. There are several line of evidence that two species of *Senecio* which occur on Mount Etna may represent an example of the early stages of ecological speciation in the face of gene flow.

Senecio aethnensis and *Senecio chrysanthemifolius* are a potential example of recent ecological speciation with gene flow. However, prior to the analyses presented in this thesis, several questions remained outstanding. For example, there was no estimates of important demographic parameters, such as their divergence time, and no formal test of the hypothesis of gene flow between the species since their divergence. These are noteworthy for several reasons. For instance, if they had diverged long before the growth of Mount Etna had begun, then a hypothesis of *in situ* divergence would clearly be wrong.

Naively, divergence time would appear to be a simple statistic to calculate if the mutation rate and rate of single nucleotide polymorphisms (SNPs) between the species were known, simply by multiplying these numbers together. However, because these species are known to hybridise, this approximation is meaningless since gene flow between the species would reduce the level of genetic divergence. Several techniques have been developed to untangle these signals (Pinho & Hey 2010). These typically rely on the variation in rates of divergence among loci, since substantial rates of gene flow are expected to increase the variation in divergence among loci. This signal can then be fitted to a simple model of the splitting of a population into two populations of unknown size, an unknown amount of time ago, followed by an unknown rate of gene flow, and these unknown parameters can be estimated using likelihood based statistics. This chapter uses such a method to elucidate the speciation process in these species using a massive and unbiased dataset.

Furthermore, while there are estimates of levels of genetic divergence between the species for small numbers of loci (Brennan et al. 2012), little is known about how this varies across the genome, and especially how this varies between synonymous and non-synonymous site within protein-coding sequences. This can give insight into the strength of selection

operating between the species (Nielsen & Yang 1998). Underpinning these analyses, as well as future work on the species, is the need for a sound bioinformatic basis for research in the species. No genome is yet available for any *Senecio* species, and at the time of the research in this chapter, no high-throughput genetic information was available for the species at all.

This chapter aims to rectify these issues with a large transcriptomic dataset of *S. aethnensis, S. chrysanthemifolius* and an outgroup *S. vernalis* with the specific aims of: i) producing the first reference transcriptomes for these species; ii) estimating the demographic parameters of their speciation (to be compared to the timescale of Mt. Etna's growth), and iii) estimating the strength and direction of selective pressures acting across the transcriptome and how this varies among loci. For this purpose, *Illumina* RNA-seq was used to sequence the transcriptomes of one individual each of the three species. In addition to evolutionary analyses, the data were used to further develop a public database (http://www.seneciodb.org/) which represents an important resource for further investigations into the evolutionary genomics of *Senecio*.

2.2 Materials and Methods

2.2.1 Sequencing

A single plant each of *S. aethnensis, S. chrysanthemifolius* and *S. vernalis* were grown in the glasshouse from seeds collected in the wild. *S. aethnensis* was collected from 37.42ºN, 14.59ºE, 2097m above sea level; *S. chrysanthemifolius* was collected from 57.57ºN, 14.57ºE, 763m above sea level (both locations are on Mount Etna, Sicily); *S. vernalis* seeds were obtained from the Millennium Seed Bank and were originally collected in Cyprus.

Total RNA was extracted from actively growing shoots with a single capitulum bud to maximise the number of genes represented in the transcriptome. Extractions were conducted using the Qiagen Plant RNeasy kit. Poly-A selection, reverse transcription, library construction and sequencing was performed according to Illumina RNAseq protocol at the genomic sequencing facility at the Wellcome Trust Centre for Human Genetics (WTCHG) in Oxford to produce 100 base pair (bp) paired-end reads. Raw data were uploaded to the Sequence Read Archive (SRA) database (Study Accession Number: SRP028289).

2.2.2 Dataset preparation

Basecalling, adaptor trimming and sorting of reads by multiplex tags were undertaken as part of the WTCHG bioinformatic pipeline. This uses Illumina's native basecalling pipeline (*Bustard* 1.9) with default parameters, and demultiplexing is performed using an in-house Perl script at the WTCHG. Reads received from the WTCHG were quality trimmed using the modified-Mott trimming algorithm in *CLC Genomics Workbench* version 5 (CLC bio, Aarhus, Denmark). Default settings of two unknown nucleotides (Ns) allowed per read and an error probability cutoff of 0.05 were used (See *CLC Genomics Workbench 5* manual for details). Amplification artefacts which occur during high-throughput library preparation can make *de novo* assemblies inaccurate (Kozarewa et al. 2009) so the *CLC Genomics Workbench Duplicate Reads Removal Plugin* was used to remove them. This uses an algorithm which distinguishes duplicate reads arising from amplification artefacts (as opposed to identical reads in high coverage regions) by first identifying 'neighbours' (reads that share most of their sequence, but with an offset). It then identifies read pairs which share identical sequence at high copy numbers compared to neighbours, and in which all, or nearly all,

duplicates are on the same strand. Identical read pairs with such a signature are extremely unlikely to occur by chance so these were reduced to one copy before assembly.

De novo assembly of the transcriptomes from each individual was conducted using *CLC Genomics Workbench* version 5. with a k-mer length (termed word size in the *CLC Genomics* documentation) of 28 bp. Other settings used were a minimal contig length of 300bp, automatically determined maximum bubble size, and use of the scaffolding algorithm, that uses paired-end information to scaffold the contigs (Using settings: mismatch cost = 3; insertion cost = 3; deletion cost = 3; similarity fraction = 0.95; length fraction = 0.5).

To investigate divergence between the species, a reference guided mapping approach was used, with reads mapped to the assembled transcriptome of *S. vernalis*, which yielded the greatest number of reads and contigs, as a reference. Quality trimmed reads (see above) from each of the three species were mapped to the reference using CLC Genomics Workbench v5 with strict settings (mismatch cost = 3; insertion cost = 3; deletion cost = 3; similarity fraction = 0.95; length fraction = 0.9). Binary Alignment Format (BAM) files for each mapping were then exported and input into the *samtools* package (Li et al. 2009) for local phasing, variant calling, filtering, and consensus calling. Several of the downstream analyses require haploid sequences, so firstly, *samtools phase* function was used to locally phase alleles. The *samtools phase* algorithm is conservative, since it phases heterozygous bases only when phase information is available within individual reads. When phase cannot be determined it outputs IUPAC ambiguity characters. BAM files for both phases of each species were then used to create a pileup file using *samtools' mpileup* function with a strict PHRED-scaled base quality cutoff of *Q*=30 (equivalent to a 1/1000 probability of incorrect base assignment). Indels were not called, as they were not relevant to any of the

downstream analyses, although the base alignment quality (BAQ) algorithm in *samtools* was used to decrease the probability of false positive SNPs arising from proximity to indels. Outputted Binary Call Format files (BCF) were then converted to variant call format (VCF) with *BCFtools view* function. VCF files, in conjunction with the consensus sequence, were used to create consensus fastq files for each species using the *samtools* helper script, *vcfutils.pl* (Li et al. 2009). Conversion to fastq format included an additional filtering step which required all bases called to be represented by at least 3 independent reads. Finally fastq files were converted to fasta format using seqtk (part of the *samtools* package) and bases which failed the depth and quality filters (represented as lowercase bases in the vcfutils.pl output) were converted to ambiguity characters (N's).

All filtered consensus fasta files were then imported into *Proseq3* (Filatov 2009) to create alignments for all reference contigs. Since only a single sequence for each species was required for downstream analyses, one phase of each species was randomly removed for each contig. The resulting tripartite alignments were then further filtered in *Proseq3*, with only contigs with at least 300bp of aligned sequence for all three species retained for further analysis.

2.2.3 Transcriptome annotation and validation

To annotate the reference *de novo* transcriptome assembly, to filter out sequences of dubious origin (i.e. sequences from contaminating organisms and plastid genomic sequence), and to assess the completeness of the transcriptome, a series of BLAST based analyses were conducted. Firstly, sequences for each locus were used as queries in BLASTX searches (Altschul et al. 1990) against the NCBI non-redundant (nr) protein database, using BLAST2GO (Conesa et al. 2005) with default settings. Gene Ontology terms (GO

Consortium) were then assigned using BLAST2GO based on the results of the BLASTX Results were manually filtered by species and those with top hits from non-plants were removed from further analysis. The resulting functional annotations were then imported and assigned to the aligned datasets using *Proseq3*. To remove plastid genome sequence from the dataset, a *BLASTN* search was performed against the closely related *Jacobea vulgaris* (formerly *Senecio jacobea*) plastid genome sequence (Doorduin et al. 2011). Sequences with over 95% identity over 100bp with *J. vulgaris* cpDNA were removed from further analysis. To assess the quality of the transcriptomes, a second BLASTX search was performed against the TAIR10 *Arabidopsis thaliana* proteins (ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/TAIR10_pep_20101214; Downloaded 05-01-12; Swarbreck et al. 2008) using the *BLAST+* suite (version: BLAST2.2.25+; Camacho et al. 2009) with default length and *E*-value settings. The length of the BLASTX alignment and that of the top-hit sequences were then compared to evaluate transcript completeness. A final filtering step based on ORF analysis was then conducted in *Proseq 3.* To minimise errors arising from incorrect coding region assignment, loci which included premature stop codons or more than one overlapping putative CDS in a different reading frame or orientation were removed. The inferred coding sequences (CDS) were used to make a coding-region only dataset for subsequent divergence analysis.

2.2.4 Testing the mode of speciation

To determine whether the species had diverged in the presence of gene flow, and to estimate their divergence times and population sizes, the likelihood ratio test and maximum likelihood models implemented in 3s were used (version 2.0a; Yang 2010). In this analysis only four-fold degenerate sites were used to minimise the influence of

selection and to allow the mutation rate to be estimated more accurately. Filtering of four-fold degenerate sites from the 'CDS only' datasets was undertaken using *Proseq3* (Filatov 2009). 3s was run with default settings with the exception that 32 points (parameter *k* in 3s) were used in the Gaussian quadrature, which produces a more accurate, but more computationally expensive, result (Yang 2010). The 3s program's implementation of the model of Yang (2002) was also used to estimate four demographic parameters of the species' divergence, namely: $\vartheta_{acv} = 4N_{acv}\mu$; $\vartheta_{ac} = 4N_{acv}\mu$; $\tau_{acv} = T_{acv}\mu$; and $\tau_{ac} = T_{ac}\mu$; where N is the effective population size, T is the divergence time (in years, since generation time was assumed to be 1 per year), μ is the mutation rate, and subscript letters denote the common ancestors of all three species (acv) and of just the two focal species (ac). These were converted into more biologically meaningful units as follows: $N_e = \vartheta/(4\mu)$ for the effective population sizes of each ancestral population; and $T = \tau/\mu$, for the divergence times to each ancestral node. The mutation rate is not known in *Senecio*, although a neutral mutation rate of 1×10^{-8} has been reported within the *Asteraceae*, of which *Senecio* is a member (used in Strasburg & Rieseberg 2008). The *Asteraceae*-specific rate was used for conversions, but because the average plant mutation rate estimated by Wolfe et al. (1987; 5×10^{-9}) has been used in a previous demographic analysis of *Senecio* (Muir et al. 2013), parameter estimates were also scaled using this mutation rate for enhanced comparability between the studies. The *Asteraceae*-specific rate is likely to be the most accurate for *Senecio*. 3s analyses were run twice to ensure that the results were stable.

2.2.5 Divergence analysis

To compare the divergence between the two Etnean species from their outgroup *S. vernalis,* genes were concatenated and subjected to Tajima's relative rate tests (Tajima

1993) implemented in *MEGA 5* (Tamura et al. 2011). Branch-specific d*N*, d*S* and d*N*/d*S*, were estimated with the *codeml* program in *PAML* version 4 (Yang 2007) using the M1, free-ratios model. To determine whether the median d*N*/d*S* from the outgroup to *S. aethnensis* and from the outgroup to *S. chrysanthemifolius* were significantly different, a Wilcoxon signed rank test was used. To detect positive selection, several pairs of nested models in codeml were compared using likelihood ratio tests (LRTs). The PAML branch site test of positive selection (Yang & Nielsen 2002; Yang et al. 2005; Zhang et al. 2005), detects positive selection at a subset of sites in a specific lineage. Two branch-site models, one specifying positive selection at a subset of sites on each of the *S. aethnensis* and *S. chrysanthemifolius* lineages (*model* = 2, *NSsites* = 2), were compared to the null model with d*N*/d*S* in tested branches set to 1. Two pairs of site models, which detect selection at a subset of sites, were also compared with LRTs (*M7* vs. *M8* and *M1a* vs. *M2a*; Nielsen & Yang 1998; Yang et al. 2000). Correction for multiple tests was implemented using the false discovery rate method (FDR; Benjamini & Hochberg 1995) using an alpha level of 0.05 to determine significance. All *PAML* analyses were automated using in-house *Perl* scripts.

2.2.6 Factors affecting the genome-wide selective landscape.

To investigate which factors may affect the differences in selective constraint among loci in the species, the d*S* and d*N* estimates for the whole tree (i.e. the sum of the estimates from all three branches estimated in codeml, above) were used to calculate d*N*/d*S* for each of the contigs. Those for which d*N*/d*S* could not be calculated (i.e. those in which d*N* or d*S* was zero) were excluded from this analysis.

Following gene duplication, selection pressures may be altered relative to non-duplicated genes, and those which tend to be retained as duplicates may preferentially belong to

functional groups which are under different selection pressures to those which are not. Therefore, to determine the relationship between d*N*/d*S* and the presence/absence of paralogues, duplicate genes within the reference transcriptome were predicted using the method of McKain et al. (2012). Genes were then divided into two groups, those with at least one paralogue in the transcriptome versus those without, and the distributions of d*N*/d*S* in two groups of genes were compared using a Mann-Whitney-U test.

To determine the level of correlation between d*N*/d*S* and the number of paralogues present, and whether this significantly differed from zero, a Spearman's rank correlation test was conducted between d*N*/d*S* for each locus and the number of paralogues detected. To investigate the possibility that any difference in d*N*/d*S* was due to inflated d*S* (either an artificial increase because of incorrect alignment of paralogues or a real one, due to selection on synonymous sites) a further Spearman's rank test was applied to d*S* and number of paralogues.

In other species (Slotte et al. 2011; Koonin & Wolf 2006) expression level has been found to be correlated with the level of selective constraint, so the relationship between expression level and d*N*/d*S* in *Senecio* was then examined. Reads per kilobase per million mapped reads (RPKM) was determined using *CLC Genomics Workbench* version 5. This was tested for correlation with d*N*/d*S* using a Spearman's rank correlation test.

To determine whether there was enrichment of any specific GO terms in genes with high d*N*/d*S*, one-tailed Fisher's exact tests were performed for each GO term in *BLAST2GO* (Conesa et al. 2005) with FDR correction for multiple tests. These were performed comparing loci in the top 5% and 10% for d*N*/d*S* to all other loci.

Genes with more distinct functions may be expected to be under stronger selective constraint. To assess the association between the level of purifying selection on a locus and the 'multifunctionality' (*sensu* Salathé et al. 2006) of the protein it encodes, total numbers of GO terms assigned to each locus, as well as the number from each of the three main GO ontologies, were compared to d*N*/d*S* using Spearman's rank correlation coefficient.

Tests for enrichment of GO terms in loci with high d*N*/d*S* and the comparisons of the distribution of GO term number with that of d*N*/d*S* could potentially be biased by some genes (e.g. especially those with low expression) being fragmented into several contigs in this dataset (and GO terms being counted multiple times for such genes). To control for this, the STM scaffolding method was used (Surget-Groba & Montoya-Burgos 2010), to scaffold contigs which may be derived from single transcripts by comparison to the *Arabidopsis thaliana* proteome (TAIR 10; see above). Contigs were subjected to a *BLASTX* search, using the suggested e-value cutoff of 10^{-5}, and xml outputs were run through the STM program using default settings. Contigs which were scaffolded using this process were then merged into single entries (and their GO terms were combined), and the tests were re-performed. The STM scaffolding method was not used for the main dataset because it introduces an increased risk of producing chimeric sequences, and is thus less conservative for most analyses.

2.3 Results

2.3.1 Transcriptome sequencing, validation and characterisation

In total ca. 1.48, 1.23 and 3.64 Gbp of sequence data were generated for *S. aethnensis, S. chrysanthemifolius* and *S. vernalis* transcriptomes, respectively. *De novo* assemblies of

these sequences yielded 30,560, 26,536, and 35,770 contigs over 300 bp respectively (with total lengths of: *S. aethnensis*: 26,897,354; *S. chrysanthemifolius*: 21,356,997; and *S. vernalis*: 27,343,943). De novo assembled contigs for these transcriptomes were uploaded to the *SenecioDB* database (http://www.seneciodb.org/).

As an approximate estimate of transcriptome completeness and quality, a further *BLASTX* search against the *Arabidopsis thaliana* proteome was performed and the proportion of the top hit which was covered by each contig in the *BLASTX* alignment was determined. This revealed that, (after filtering described in methods) 23.42% of *Senecio* contigs covered at least 90% of their top hit, suggesting that these sequences may represent approximately full length transcripts. 57.09% of contigs covered at least 50% of their top hit. The contigs were annotated with a wide range of GO terms, suggesting that a diverse array of functional gene categories were captured (Osborne et al. 2013 - supplementary data). After filtering, 12,679 contigs were successfully assigned coding regions longer than 100 codons and these made up the CDS-only dataset for subsequent d*N*/d*S* based analyses.

To estimate the number of contigs which may have been disconnected regions of the same transcript, an STM scaffolding approach that takes advantage of homology and functional annotation to scaffold cDNA contigs likely to come from the same gene was used (Surget-Groba and Montoya-Burgos, 2010). This analysis identified only 498 (3.93%) contigs which were putatively fragments of a larger transcript and were assembled by *STM* into 214 scaffolds. While this approach can increase the contiguity of a transcriptome, it also risks creating chimeric sequences from fragments of related but distinct genes. Since it relies on a high quality, closely related reference proteome, which is not available for *Senecio* or close relatives (*Arabidopsis thaliana* was used as the reference in this analysis), this risk is

greatly increased, and could severely affect the d*N*/d*S* and demographic analyses. For this reason, it was more conservative to use the non-STM scaffolded reference for the majority of these analyses. The scaffolded transcriptome was only used as a control for GO enrichment analyses.

In the analysis of paralogue evolution, 2,598 genes were identified as having a putative paralogue among the whole reference assembly (20.49% of the transcriptome). The distribution of the number of paralogues present (Osborne et al. 2013 - supplementary data) revealed that over 75% of identified genes with paralogues had only one transcribed paralogue. Median synonymous divergence (d*S*) between paralogous pairs was 0.90, and paralogues were significantly enriched for 190 GO terms (when reduced to most specific terms; Osborne et al. 2013 - supplementary data). Using the *STM* dataset for the same analysis the results were broadly similar (187 GO terms were over-represented in the *STM* dataset all together, and 175 GO terms were significantly over-represented in both datasets; Osborne et al. 2013 – supplementary data). These results suggest that there is a high proportion of retained paralogues in *Senecio aethnensis* and *S. chrysanthemifolius* and that many paralogues are expressed simultaneously.

2.3.2 Testing the mode of speciation

S. aethnensis and *S. chrysanthemifolius* might be assumed to represent a good example of ecological speciation (Chapman et al. 2005; James & Abbott 2005; Brennan et al. 2009) as these species are closely related (and likely sister species), but adapted to contrasting environments. Indeed, mean synonymous (d*S*) and non-synonymous (d*N*) divergence between the orthologous genes of the two Mt. Etna species was quite low (d*S* = 0.016 ± 0.017 [SD]; d*N* = 0.002 ± 0.003[SD]), which is consistent with a recent origin of these

species. To determine whether substitution rate differed between the species, a Tajima's relative rate test was performed (Tajima 1993) on concatenated alignments, however there was no significant difference between the rates of *S. aethnensis* and *S. chrysanthemifolius* (χ^2 = 1.98, *P* = 0.16). The divergence of the Mt Etna species from the outgroup *S. vernalis* was about twofold higher: (*S aethnensis*: d*S* = 0.035 ± 0.023 [SD]; d*N* = 0.005 ± 0.005 [SD]; *S chrysanthemifolius*: d*S* = 0.035 ± 0.023 [SD]; d*N* = 0.005 ± 0.005 [SD]), confirming previous findings (Comes & Abbott 2001) that *S. aethnensis* and *S. chrysanthemifolius* form a clade relative to *S. vernalis* and thus validating its choice as an appropriate outgroup for the study (Fig. 2.1).

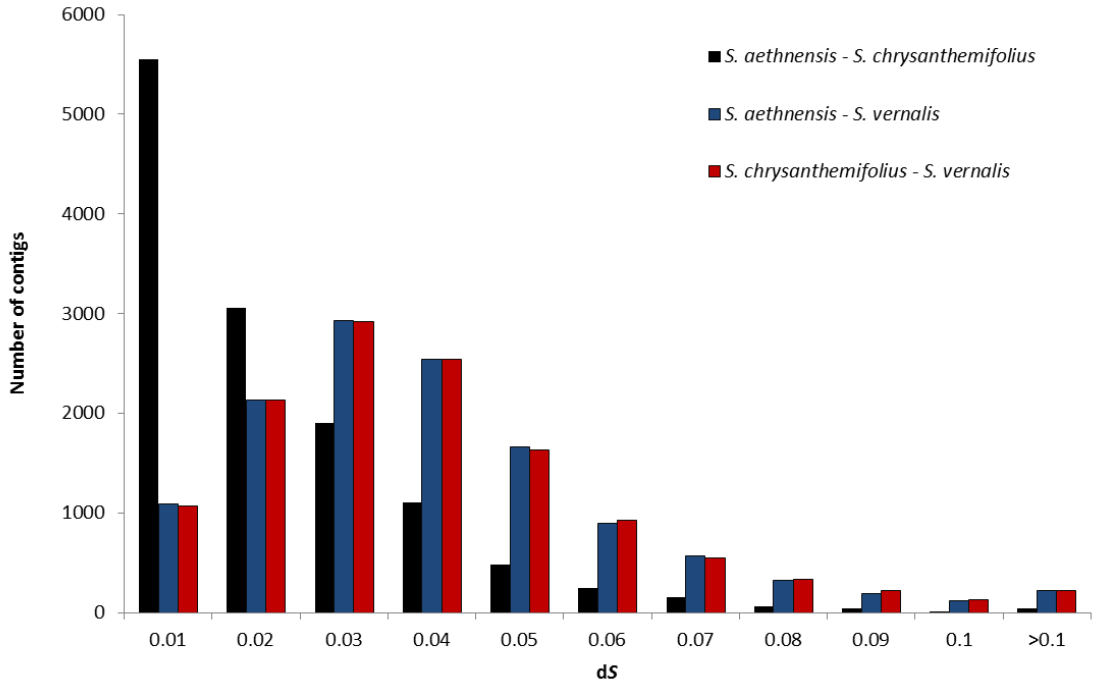

Figure 2.1: Distribution of contig-specific synonymous divergence between *S. aethnensis* and *S. chrysanthemifolius* (black), between *S. aethnensis* and *S. vernalis* (red), and between *S. chrysanthemifolius* and *S. vernalis* (blue). The x axis values are the maximum value in each bin range.

Little is known about the demographic history of *S. aethnensis* and *S. chrysanthemifolius'* speciation, and how it relates to the geological evolution of Mt. Etna. To estimate the demographic parameters of the species split the model developed by Yang (2002) was

used. This model provided raw maximum likelihood estimates (MLEs) for all four parameters (θ_{acv} = 0.0097 ± 0.0003[SE], θ_{ac} = 0.0012 ± 0.0004[SE]; τ_{acv} = 0.0108 ± 0.0001[SE]; and τ_{ac} = 0.0015 ± 0.0001[SE]; see method for parameter details). The conversion of these into more intuitively obvious demographic units requires knowledge of the mutation rate, which is not known in *Senecio.* Therefore an estimate from the *Asteraceae* (1×10^{-8}; used in Strasburg & Rieseberg 2008), was used, which is expected to be fairly accurate for *Senecio,* as well as the estimated average plant rate (5×10^{-9}; Wolfe et al. 1987) for comparability with a previous study which used it (Muir et al. 2013) that is nevertheless expected to be less accurate because of its phylogenetic generality. This was translated into ancestral population sizes of N_{acv} = 243,238 ± 6,338 [SE] for the ancestor of all three species and N_{ac} = 312,023 ± 9,833 [SE] for the ancestor of the two Etnean species using the *Asteraceae*-specific mutation rate. Using the generic plant mutation rate, population size estimates were higher (N_{acv} = 486,475 ± 12,675 [SE] and N_{ac} = 624,045 ± 19,665 [SE]). Divergence times were estimated to be T_{acv} = 1,077,760 ± 12,240 [SE] for the split between the ancestors of *S. vernalis* and the Etnean species and T_{ac} = 153,080 ± 11,470 [SE] for the split between *S. aethnensis* and *S. chrysanthemifolius*, using the *Asteraceae*-specific mutation rate. These estimates of the divergence time between *S. aethnensis* and *S. chrysanthemifolius* occur around the period in its geological history during which Mount Etna began to exceed the altitude above which *S. aethnensis* is now found, and *S. chrysanthemifolius* is not (De Beni et al., 2011; Fig.2.2.), suggesting that the creation of a new niche as the mountain grew may have been a catalyst for the plant's speciation. Divergence time estimates increased moderately when the plant mutation rate was used (T_{acv} = 2,155,520 ± 24,480 [SE] and T_{ac} = 306,160 ± 22,940 [SE]).

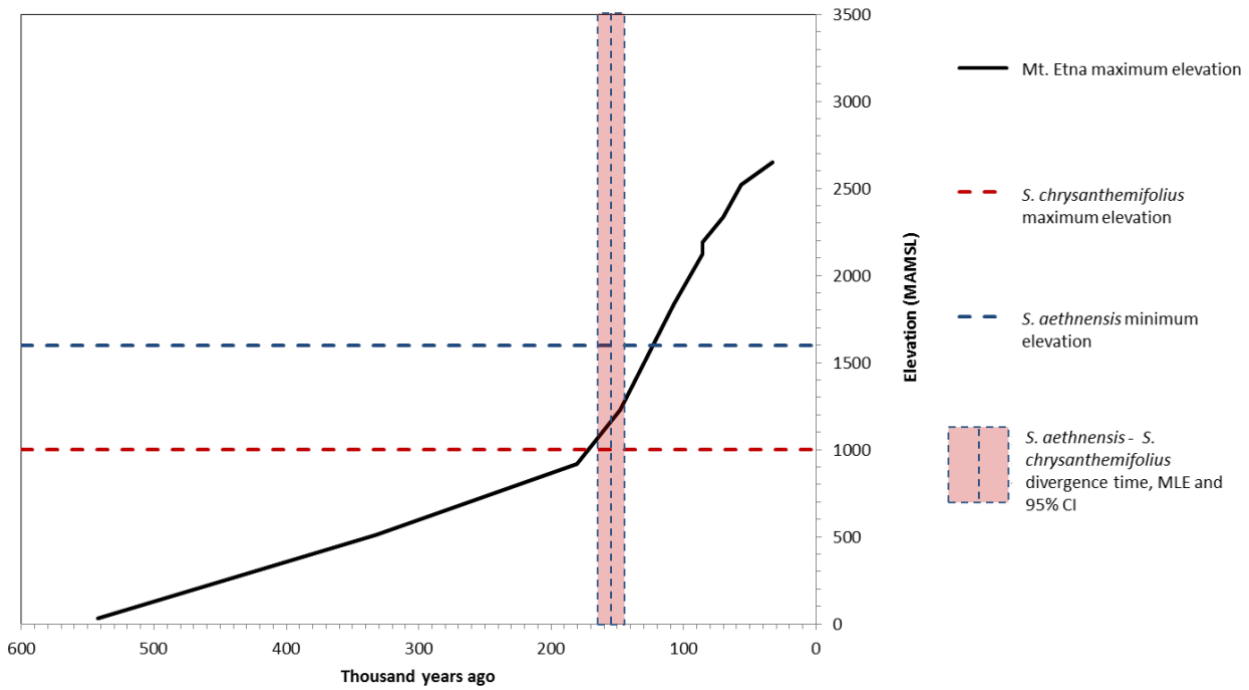

Figure 2.2: The estimated divergence time of the *S. aethnensis* and *S. chrysanthemifolius* occurs shortly after Mt. Etna exceeded the elevations that partition the species today (converted into years using an *Asteraceae*-specific mutation rate of 1×10^{-8} and a generation time of 1 year). Age of volcanic layers of different elevations estimated from 40Ar/39Ar isotopic dating in De Beni et al. (2011).

To test a null hypothesis of allopatric speciation with no significant subsequent gene flow, compared to a model of divergence with gene flow, the likelihood ratio test developed by Yang (2010) was used. The test was highly significant ($\chi^2 = 129.43$; $P = 2.73 \times 10^{-30}$), indicating that there has been significant gene flow between the species since their divergence. This is consistent with an ecological speciation model, but is also compatible with a scenario of secondary contact following a period of allopatric divergence.

2.3.3 Transcriptome-wide analyses of d*N*/d*S*

d*N*/d*S*, an indicator of the strength and type of selection, was then characterised across the transcriptome; and its relationship to several important parameters: expression level, presence of paralogues and gene function was determined. Median d*N*/d*S* was 0.153 between *S. vernalis* and *S. aethnensis* and 0.148 between *S. vernalis* and *S. chrysanthemifolius*, suggesting that purifying selection is the dominant selective force acting on the loci. This ratio varied widely among the genes (Fig. 2.3) and there was only a modest significant correlation between d*N*/d*S* on the *S. aethnensis* and *S. chrysanthemifolius* branches (Spearman's rank: r_s = 0.241, $P = 2.0 \times 10^{-21}$). This correlation rises to 0.81 (Spearman's rank: $r_s = 0.810$, $P < 1 \times 10^{-100}$) when the whole distance between each of the Etnean species and *S. vernalis* is considered, but most SNPs in the dataset are between *S. vernalis* and the Etnean species, and are invariant between *S. aethnensis* and *S. chrysanthemifolius,* so this figure is misleadingly inflated. Despite, the fact that the correlation between d*N*/d*S* in the two Etnean lineages was fairly weak, there was no significant difference between median d*N*/d*S* on the *S. aethnensis* and *S. chrysanthemifolius* branches (Wilcoxon test: $W = 5.79 \times 10^5$; $P = 0.189$), indicating that the overall level of selective constraint in the two lineages is comparable.

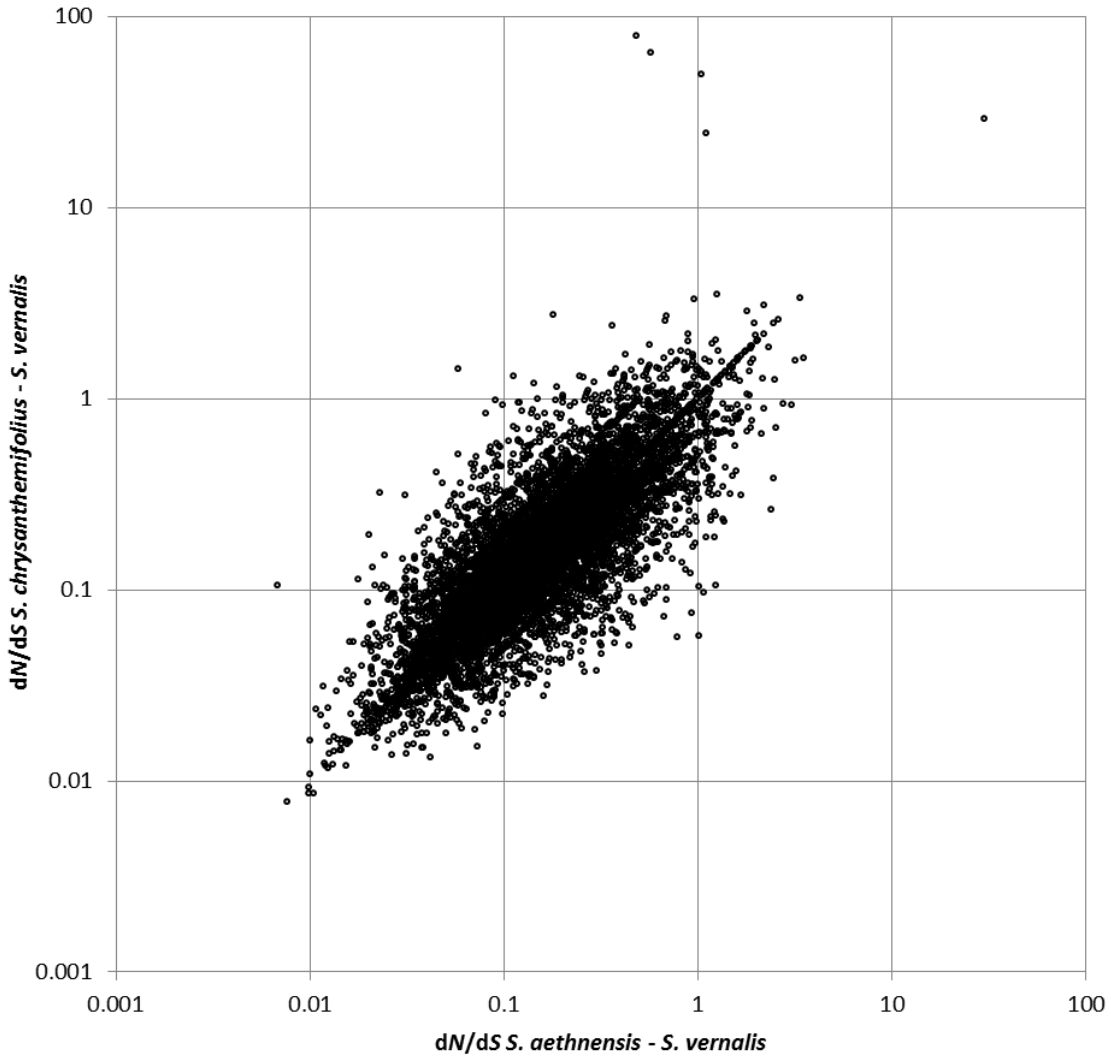

Figure 2.3: A scatterplot of d*N*/d*S* for each locus between *S. aethnensis* and *S. vernalis* versus d*N*/d*S* between *S. chrysanthemifolius* and *S. vernalis.*

Transcriptome-wide trends which may influence d*N*/d*S* were then investigated. Interestingly, across all loci, there was a weak but significant negative correlation between total level of expression and d*N*/d*S* (Spearman's rank: r_s = -0.222, $P = 1.9 \times 10^{-119}$), suggesting that more highly expressed genes are under stronger purifying selection. d*N*/d*S* was significantly lower in genes with putative paralogues (median = 0.169 ± 0.188[SD]) than those without (median = 0.237 ± 0.415[SD]; Fig. 2.4; Mann-Whitney $U = 1.18 \times 10^6$; $P = 2.97 \times 10^{-8}$; genes for which d*N*/d*S* could not be calculated were excluded from the test). Furthermore, there was a weak but highly significant negative correlation between the

number of paralogues for a gene and its d*N*/d*S* ratio compared to orthologues (Spearman's rank: r_s = -0.131, $P = 4.33 \times 10^{-42}$), suggesting that genes which are retained as paralogues, tend to be under stronger purifying selection. It is possible that a correlation between number of paralogues and d*N*/d*S* could arise because, in genes which are part of high copy gene families, there could be an increase in incorrect mapping of reads from their many different paralogues. This would also produce an increase in d*S* in high copy number gene families. There was no significant correlation, however, between d*S* and number of paralogues, suggesting that this was not the case.

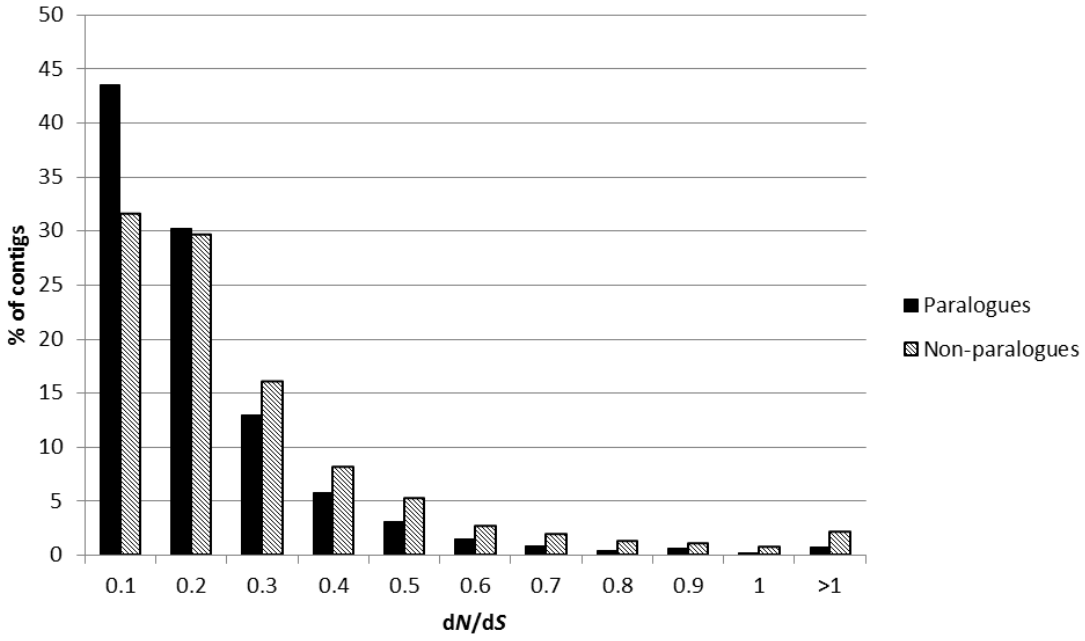

Figure 2.4: Genes with expressed paralogues (black) in the transcriptome have significantly lower d*N*/d*S* than those without (grey). The x axis values are the maximum value of d*N*/d*S* in each bin range. The percentage of contigs (in each of the two classes, i.e., with or without paralogues) in each bin are displayed on the y axis.

Finally the relationship between predicted function and d*N*/d*S* was examined. Fisher's exact tests with FDR correction showed that genes with the highest d*N*/d*S* were not significantly enriched for any GO terms (neither genes in the top 5% nor 10% tails of the d*N*/d*S* distribution). This was true when d*N*/d*S* in the *S. aethnensis* or *S. chrysanthemifolius*

branches as well as across the whole tree, was considered, and was also true of the STM-scaffolded, control dataset (see Materials and Methods). However, the number of GO terms annotated to a gene, ('multifunctionality'; *sensu* Salathé et al. 2006), showed a highly significant but weak negative correlation with d*N*/d*S* (Spearman's rank: r_s = -0.191, P =1.31 × 10^{-88}). This was also true when GO terms were divided into their three ontology categories; molecular function (Spearman's rank: r_s = -0.127, P = 7.02 × 10^{-40}), cellular component (Spearman's rank: r_s = -0.169, P = 2.17 × 10^{-69}), and biological process (Spearman's rank: r_s = -0.152, P = 1.58 × 10^{-56}). This may be evidence that more multifunctional genes, those which are involved in a greater number of distinct processes at molecular or organismal levels, or that function in more cellular locations, are more likely to experience a high level of purifying selection.

To attempt to identify specific loci under positive selection several d*N*/d*S* -based tests of positive selection were used (Nielsen & Yang 1998; Yang et al. 2000; Yang & Nielsen 2002; Yang et al. 2005; Zhang et al. 2005). Using an FDR cutoff of 0.05, however, no locus showed evidence for positive selection in the branches leading to either *S. aethnensis*, *S. chrysanthemifolius*, or across the whole tree. However, with only three species and a very large number of loci, the test may have lacked sufficient power.

2.4 Discussion

2.4.1 Characterisation of the *Senecio* transcriptomes.

Genome or transcriptome-wide datasets for closely related species are essential for evolutionary genomic analyses, but they have historically been restricted to a few model species. However, with the advent of next generation sequencing, the situation has begun

to change, with genomic resources for non-model organisms, becoming available (Rokas & Abbot 2009). *De novo* sequencing and assembly of non-model organisms, however, presents significant bioinformatic challenges, particularly if, as in *Senecio*, no high quality reference genome or transcriptome is available from a closely related species. Transcriptome sequences for three *Senecio* species generated in this chapter represent the first step towards the development of comprehensive genomic resources for this fascinating system, which promises to be a rich resource for studies of adaptation and speciation in plants. *De novo* assembly and annotation of cDNA contigs, over 20% of which covered the full length of their nearest Arabidopsis homologue, indicate that Illumina sequencing of transcriptomes is an effective strategy for producing large and high quality comparative-transcriptomic datasets, even in the absence of a reference genome. The annotated transcriptomes presented here are the first published transcriptomes of the species and greatly increase the amount of data available to the community. *SenecioDB* (http://www.seneciodb.org/) now contains sequences from six *Senecio* species, complementing the existing Compositae Genome Project database (http://compgenomics.ucdavis.edu/). The data will thus further facilitate large scale comparative genomic analyses of divergence in one of the largest families of flowering plants (Panero & Funk 2008).

2.4.2 Recent speciation may have been driven by the growth of Mount Etna

The data were used to gain a better understanding of the demographic history of the species. Species adapting to altitude are attractive systems for the study of ecological speciation for several reasons. Firstly, four distinct environmental variables which can exert strong selective pressures (atmospheric pressure, temperature, total solar radiation and

fraction of UV-B radiation; Körner 2007) co-vary with altitude universally. Secondly, several other factors such as precipitation, seasonality and biotic interactions are often linked to elevation on a case-by-case basis (Körner 2007; Defossez et al. 2011). Thus, altitudinal gradients can exert strong divergent selection over a short geographical distance, and a simple measure of altitude can be a proxy for many of these. Furthermore, Etnean *Senecio* are unusual among cases of altitude-related speciation in that the mountain on which the species are found is a volcano, and arose relatively recently and rapidly (De Beni et al. 2011; Branca et al. 2011; Fig. 2.2). While volcanic activity in the area began over 1.5 mya, the oldest dated rock samples over 1000 metres above mean sea level (MAMSL) were formed only 147.7 kya, and the oldest over 1600 MAMSL only 107.2 kya, with the current mountain standing at over 3000 MAMSL (De Beni et al. 2011; Fig. 2.2). The finding that the species split between ~141 and ~164 kya, fits strikingly well with the time that the volcano attained the elevations which partition the species today. This suggests that the species may have diverged in response to a new high altitude niche rapidly becoming available as the volcano grew. While the estimate is dependent on the mutation rate used for conversion (Yang, 2010), two different mutation rates were used, one a general estimate for plants and the other an *Asteraceae*-specific estimate, and both give estimates of divergence time within the volcano's period of rapid recent growth. Since mutation rate is taxonomically correlated in plants (Eyre-Walker & Gaut 1997), the *Asteraceae* specific estimate is expected to be more accurate for *Senecio*.

A recent study into signatures of selection on differentially expressed versus non-differentially expressed genes in *Senecio aethnensis* and *S. chrysanthemifolius* (Muir et al. 2013) used an isolation-migration model to estimate divergence times, population size and

migration rate between the species. This found an even more recent estimate of divergence time (15-75 kya). However, (as was acknowledged by the authors) in that paper only a small number of loci were used (15 nuclear genes and 5 microsatellites) without distinguishing between nonsynonymous, synonymous and non-coding sites, and used only one mutation rate (5×10^{-9}) in conversions. Thus, their estimates are likely to be fairly approximate. The use, in the current study, of only four-fold degenerate sites means substitution rates are likely to be far more homogenous between sites, and any influence from selection is likely to be minimised. The far larger dataset used here and use of two different mutation rates in conversions resulted in more precise estimates with narrow confidence intervals.

The test of a model of divergence-with-gene-flow versus a null hypothesis of allopatric speciation was extremely significant, suggesting that gene flow has occurred since the species split. Therefore, in addition to gene flow from the parent species into the hybrid zone, there has also been significant exchange of genes between the native ranges of each species where the parents appear 'phenotypically pure' (James & Abbott 2005). The significant result cannot, however, be taken as confirmation of speciation with gene flow *sensu stricto.* This is because the test does not distinguish between divergence in the presence of gene flow and allopatric speciation followed by secondary contact and gene flow. However, the high significance of the test rejecting a no gene flow model in this chapter suggests that gene flow has occurred between these two species over sufficient time to have had a considerable impact upon the pattern of species divergence across the genome. Moreover, there are several pieces of 'circumstantial evidence' that the species diverged with gene flow. The species are wind dispersed; they display minimal intrinsic

barriers to interspecific hybridisation (none were found in Chapman et al. 2005, but some evidence for a modest level of intrinsic reproductive isolation was found in Brennan et al. 2014 and Chapman et al. 2016); they occur in close proximity to each other with no geographic barriers between them; and, presuming they diverged *in situ*, this is likely to have been the case throughout their history (Branca et al. 2011). Taking these factors into account, speciation with gene flow, in the absence of a significant allopatric phase seems the most parsimonious account of their history. Interspecific gene flow can play contrasting roles in evolution. While it can act in opposition to local adaptation, by homogenising genomic regions under diversifying selection between two species, it can also allow globally adaptive alleles to be shared between interfertile species, aiding adaptation (Seehausen 2004; Abbott et al. 2013). Genomes of such hybridising species may be expected to exhibit a mosaic structure, with high levels of divergence restricted to 'speciation islands' which contain the loci responsible for species differences (Wu 2001). Thus, the finding of divergence with gene flow between the species indicates significant gene sharing between Etnean *Senecio* species; loci containing fixed differences that may be identified in future DNA polymorphism analyses are likely to be under diversifying selection.

Taking the demographic analyses together, in addition to previous (and subsequent) work on the system (Chapman et al. 2005; Brennan et al. 2009; Muir et al. 2013; Chapman et al. 2013) and on the geological evolution of Mount Etna (De Beni et al. 2011; Branca et al. 2011) the most plausible scenario of the species divergence is as follows. Several hundred kya, Mount Etna rapidly began growing in altitude, with the oldest dated sections above which pure *S. chrysanthemifolius* does not grow today being formed around 148 kya and

those above which *S. aethnensis* is found today around 108 kya (De Beni et al. 2011). This led to the creation of a new high altitude niche and the very different environments of the mountain's upper and lower slopes created strong divergent selection between the plants growing in the two habitats and they subsequently diverged, although gene flow persisted. Thus, the species appear to be a classic example of ecological speciation in response to rapid geological upheaval.

2.4.3 The genome wide landscape of selection in *Senecio*

Since selection is likely to have played a major role in the species' divergence, the selective landscape across the genomes of the two species was then investigated. d*N*/d*S* varied widely across the genome but was correlated in the two Etnean lineages, although several genes had much higher d*N*/d*S* in one species than the other on visual inspection of the data (Fig. 2.3). The fact that no genes had a significant signature of selection at an FDR cutoff of 0.05 is most likely due to a lack of power. Firstly, *PAML* likelihood ratio tests (Yang 2007) lack power with few species and secondly, the false discovery rate (FDR) over several thousand tests requires an extremely high level of significance.

More revealing was the investigation of d*N*/d*S* in relation to other genomic parameters. Recent work in the emerging field of evolutionary systems biology, has begun to uncover several 'laws of genome evolution' (Koonin & Wolf 2006; Koonin 2011), such as genome-wide correlations with evolutionary rate. Several studies have shown that gene expression level is negatively correlated with d*N*/d*S* (Slotte et al. 2011; Koonin & Wolf 2006) potentially due to increased selection for translational robustness against protein mis-folding in highly expressed genes (Drummond et al. 2005). The results are consistent with these previous studies, finding a negative correlation between d*N*/d*S* and expression level.

A significantly lower d*N*/d*S* in genes with paralogues, compared to those without paralogues was perhaps more surprising.

Gene duplication and subsequent functional divergence is thought to be one of the most important mechanisms for the generation of evolutionary novelty in plants (Ohno 1970; Moore & Purugganan 2005) and recently duplicated genes are expected to have higher d*N*/d*S* ratios. This is based on the prediction that immediately following gene duplication, functional constraint is reduced and therefore d*N*/d*S* might be higher as one or both paralogues evolve complementary (or novel) functions, or that one paralogue becomes pseudogenised. However, in contrast to this theory, genes in the transcriptomes used here that had at least one paralogue had a significantly lower d*N*/d*S* than those genes with no paralogues, suggesting they were under stronger purifying selection. One explanation for this finding comes from the observation that the major peak in the d*S* distribution (between paralogues) was around 0.9-1.0 (Osborne et al 2013 - Supplementary data) likely corresponding to the whole genome duplication (WGD) detected at the base of the *Asteraceae* ~50 mya by Barker et al. (2008). Thus, any relaxation in purifying selection immediately following WGD may long have passed. Indeed, it is possible that, following a WGD, many of the genes which are retained as pairs of paralogues (as opposed to one copy undergoing pseudogenisation) may preferentially be from functional categories which are particularly highly constrained. Several studies have shown evidence for two such opposing forces affecting the value of d*N*/d*S* in duplicate genes (Davis & Petrov 2004; Jordan et al. 2004; Yang & Gaut 2011). While there is a decrease in purifying selection immediately following duplication, which may lead to a higher d*N*/d*S*, there is also a general tendency for genes which are retained in duplicate to be more conserved over longer timescales,

leading to a decreased d*N*/d*S* (Jordan et al. 2004). Hence, while a higher than average d*N*/d*S* ratio (caused by relaxed purifying selection) may be expected in recent duplicates, over much longer timescales after whole genome duplications, the opposite may be observed, and the trends may be different for different functional categories of genes such as slowly evolving transcription factors or rapidly evolving immune response genes.

While there were no significantly over-represented GO terms in those loci with high values of d*N*/d*S*, the overall number of GO terms was negatively correlated with d*N*/d*S*. Genes with a greater number of distinct functions may be more constrained because they are more pleiotropic (Salathé et al. 2006), and are thus more likely to disrupt a phenotype deleteriously. However, it should be noted that these GO annotations were ascertained by sequence homology rather than direct experimentation so should be taken with caution. Nevertheless, the fact that a negative correlation still exists even with the poor functional annotation available for *Senecio,* may indicate that, with more accurate annotation, multifunctionality may be a reliable predictor of the strength of purifying selection on a locus.

The correlations of d*N*/d*S* with duplication status, expression level and multifunctionality, which the analyses uncovered, add to the growing body of evidence that the evolutionary rates of genes and proteins are influenced by surprisingly constant relationships with several aspects of gene function (Koonin 2011; Yang & Gaut 2011). Here it is shown that, even in species which appear to be undergoing strong diversifying selection, and over the relatively short time since the species diverged, many of these 'laws' appear to hold. Thus, while a subset of the genome may be involved in adaptive responses to the species'

radically different habitats, the majority of loci may continue to evolve at a rate overwhelmingly determined by variation in the strength of long-term purifying selection.

2.4.4 Conclusions

Taken together, the results in this chapter suggest that, in spite of their considerable phenotypic divergence, the two focal species, *S. aethnensis* and *S. chrysanthemifolius*, are extremely closely related genetically and that there has been a substantial amount of gene flow since their divergence. Intriguingly, the estimated time of divergence is strikingly close to estimates of the time when the height of Mount Etna was first approaching the altitude above which *S. aethnensis* occurs today, alluding to the possibility that the volcano's rapid emergence as a mountain of over 3000 MAMSL could have been responsible for the species' divergence.

The analyses in this chapter were not primarily designed to detect specific loci under selection, but it does undertake the first characterisation of selective constraint across the genome. Identification of specific loci under divergent selection and the extent to which gene flow varies between loci will allow a far greater understanding of the genomic basis for such dramatic phenotypic divergence, in the face of high levels of gene flow in such a short space of time and this should be a key focus of future work.

Chapter 3: Maintenance of species boundaries despite ongoing gene flow in Mediterranean *Senecio.*

Preface

This chapter has been published in a slightly different format as “Maintenance of species boundaries despite ongoing gene flow in ragworts“ in the journal *Genome Biology and Evolution* (volume 8, issue 4, pp. 1038-1047). All experimental work and writing is my own with the following exceptions. Dmitry Filatov, Mark Chapman, Bruno Nevado, Lynsey Bunnefeld, Konrad Lohse, and three anonymous reviewers provided comments and corrections on the manuscript. Mark Chapman grew the plants used and extracted RNA for sequencing. Bruno Nevado provided scripts for conversion of variant calls to fasta format, and concatenation of fasta alignment files.

3.1 Introduction

The “tree of life” has been one of the most enduring metaphors in evolutionary biology. The sole illustration in “On The Origin…” (Darwin 1859) depicts a model of species diversification in which speciation is bifurcating and irreversible. While Darwin and Wallace’s penetrating insights thrust divergence by natural selection to the fore (Kunte & Agashe, 2015; Darwin 1859; Pinho & Hey 2010), the homogenising effect that gene flow could play in preventing divergence was emphasised later. Many biologists have maintained that an external barrier to gene flow is necessary for speciation to occur (Dobzhansky 1935; Mayr 1963). It has long been known, however, that speciation without gene flow and tree-like evolution is an incomplete explanation of species diversification. Introgression and incomplete lineage sorting result in different phylogenetic histories for different regions of a species’ genome (Pamilo & Nei 1988) and the hybrid origin of some taxa create reticulate nodes in the tree of life (Rieseberg 2006). Furthermore, it is becoming increasingly clear that species divergence can proceed without an initial external barrier to gene flow, and divergent selection may be sufficient to drive the process of divergence, with reproductive isolation coming much later (Rundle & Nosil 2005).

Interspecific gene flow and introgression have been detected in plants (Muir et al. 2012; Strasburg & Rieseberg 2008; Arnold et al. 1992; Papadopulos et al. 2013), animals (Nevado et al. 2011; Fontaine et al. 2014; Pardo-Diaz et al. 2012) and many other groups of sexual organisms (Neafsey et al. 2010; Sun et al. 2012; Zardi et al. 2011). Introgression can have a detrimental impact on the species involved, by homogenising the regions of their genomes which have become divergently locally adapted, potentially leading to despeciation or the extinction of one taxon (Webb et al. 2011; Rhymer & Simberloff 1996). Conversely, it can

be adaptive, by increasing variation within, and sharing globally adaptive mutations between, species and creating novel combinations of alleles (Seehausen 2004). What is clear is that interspecific gene flow is clearly prevalent in a wide range of taxa in which species integrity has not been completely compromised (Whitney et al. 2010; Twyford & Ennos 2012). Nevertheless, there are relatively few examples of studies investigating multiple species within a taxon, particularly with the high-throughput datasets needed to accurately represent the complex variation in phylogenetic signal which may be present throughout the genome.

The gene flow between *Senecio aethnensis* and *S. chrysanthemifolius* uncovered in Chapter 2 of this thesis (Osborne et al. 2013), and corroborated in further work (Chapman et al. 2013; Muir et al. 2013; Filatov et al. 2016) begs the question of whether these two species represent an oddity within the genus, or if interspecific gene flow has been a common factor during the evolution of the clade. Aside from *S. aethnensis* and *S. chrysanthemifolius,* several cases of hybrid speciation have been documented in the Mediterranean *Senecio* species complex (*sensu* Comes & Abbott 2001; Kadereit et al. 2006; Lowe & Abbott 1996, 2004; Pelser et al. 2012), suggesting that some members of the clade are capable of hybridising, or at least have been in the past. Nevertheless, singular instances of hybrid speciation may represent rare events, and reproductive isolation between hybrids and their parent species may still lead to the absence of introgression between species (Rieseberg 2006). A handful of authors have documented introgression between members of the clade other than *S. aethnensis* and *S. chrysanthemifolius*. For example, a form of the normally selfing *S. vulgaris*, *S. vulgaris* var. *hibernicus* which had previously lost the ability to form ray florets (flowers on the periphery of the inflorescence, with large petals which

are attractive to pollinators) has recently arisen, which has ray florets and a higher instance of outcrossing (Abbott et al. 1992). The British endemic *S. squalidus*, itself the product of homoploid hybrid speciation between *S. aethnensis* and *S. chrysanthemifolius*, has been found to be the source of introgressed genes conferring ray florets in these plants (Abbott et al. 1992; Kim et al. 2008). Comes & Abbott (1999) found evidence of introgression of plastid DNA between *S glaucus* and *S. vernalis* in a contact zone in Israel. However, there was evidence for a far lower rate of nuclear introgression, possibly as a result of selection against immigrant nuclear DNA (Comes & Abbott 1999). Finally, in their survey of the clade, Comes & Abbott (2001), found evidence of introgression between *S. rupestris* and *S. vernalis* in parapatric populations in Greece, although they discounted introgression between *S. gallicus* and two other diploids species - *S. leucanthemifolius* and *S. glaucus.* Similarly, Coleman & Abbott (2003), and Chapman & Abbott (2005) failed to find evidence of nuclear introgression between *S. leucanthemifolius* and *S. glaucus,* and *S. gallicus* and *S. glaucus* respectively. One respect in which these previous studies are similar is that they use relatively small molecular datasets, consisting of ITS or plastid sequences or allozyme, RAPD or RFLP genotypes. None use the high-throughput sequencing data which is now available, and which produces more data by several orders of magnitude. This is significant, because high-throughput sequence data is expected to provide far greater power for detecting gene flow (Luikart et al. 2003; Twyford & Ennos 2012) and allows powerful, genome-wide tests of introgression (Durand et al. 2011). The lack of previous high-throughput data has also likely stymied attempts to derive a robust phylogenetic hypothesis for the clade (Comes & Abbott 2001; Pelser et al. 2011). Determining the phylogenetic relationships of a study system is crucial for understanding a multitude of aspects of its evolution, so this is also a priority for the group.

This chapter aims to address these issues with a genome-wide dataset covering several members of the clade. Using RNA-seq data, i) the species-level evolutionary history of group was estimated, ii) The extent of gene tree-species tree incongruence, which may have complicated previous phylogenetic and taxonomic analyses, was investigated, and iii) past introgressive hybridisation amongst the species was detected including several instances which were previously unknown. The results shed light on the process of species diversification with gene flow and suggest that introgression has occurred with surprising frequency in the clade.

3.2 Materials and Methods

3.2.1 Seed collection, plant growth and sequencing

Six species were sampled from the Mediterranean *Senecio* species complex, *S. aethnensis, S. chrysanthemifolius, S. leucanthemifolius, S. glaucus, S. gallicus* and *S. vernalis.* In addition, two outgroups species, *S. madagascariensis* and *S. flavus* were used to ensure that any effects of past introgression with the outgroup on downstream analyses of introgression were minimised. Plants were grown from wild-collected seed (locations shown in Table 3.1). Seeds were germinated on damp filter paper and seedlings were transferred to a soil/vermiculite mix in a growth room set at 19–21˚C with a 16 hour photoperiod. To maximise the number of transcripts present, apical tissues were harvested from each plant (inflorescence, stem and first apical leaf) when the first inflorescence opened, and frozen in liquid nitrogen. Tissue samples were ground whilst frozen and RNA was extracted with a Qiagen RNeasy plant kit (Qiagen, Crawley, UK) according to manufacturer's instructions, and included an optional treatment with DNase (Qiagen). Three µg of RNA per specimen was sent to the Wellcome Trust Centre for Human

Genomics, Oxford (WTCHG) for sequencing. RNA paired-end libraries were prepared individually and then combined prior to sequencing using the Illumina Hiseq 2000 sequencing platform.

Table 3.1: Locations of sampled plants and number of reads obtained for each species.

Species	Location	Number of reads
Senecio aethnensis	Mount Etna, Sicily, Italy	15,154,686
Senecio chrysanthemifolius	Randozzo, Sicily, Italy	12,591,356
Senecio flavus	Puerto de la Peña, Fuerteventura, Canary Islands, Spain	32,262,295
Senecio gallicus	Amoreira, Algarve, Portugal	19,038,966
Senecio glaucus	Morocco	31,234,638
Senecio leucanthemifolius	Propriano, Corsica, France	33,151,358
Senecio madagascariensis	Kwazulu-Natal, South Africa	34,583,848
Senecio vernalis	Cyprus	36,420,882

3.2.2 Dataset preparation

Base calling, adaptor trimming and de-multiplexing of reads was undertaken as part of the WTCHG bioinformatics pipeline. This uses the native Illumina basecalling pipeline (*Bustard* 1.9) with default parameters. Raw reads for the *S. aethnensis, S. chrysanthemifolius* and *S. vernalis* individuals used were the same as in Chapter 2, and have already been published (Osborne et al. 2013) and deposited in the Short Read Archive (SRA) under the accession SRP028289. *S. leucanthemifolius, S. gallicus, S. glaucus, S. flavus,* and *S. madagascariensis* were newly sequenced. Raw reads were imported into *CLC Genomics Workbench 7* (CLC bio, Aarhus, Denmark; hereafter *CLC*). The assembly and mapping pipeline from Chapter 2 was broadly followed (Osborne et al. 2013). Reads were quality trimmed using an error probability cutoff parameter of 0.05 and a maximum of two ambiguous bases per read using the Modified Mott trimming algorithm in *CLC* (See *CLC* manual for details). Duplicate reads resulting from PCR errors were removed using the *CLC* Duplicate Read Removal

Plugin (read numbers are reported in table 4.1). *De novo* assemblies of the two outgroup species (*S. madagascariensis* and *S. flavus*) were performed separately in *CLC*. Optimal k-mer length was calculated by CLC, which was 23 bp for both outgroup species. Further settings used were a minimum contig length of 300 bp, automatically determined maximum bubble size and scaffolding using paired end information (with mismatch cost of 2, insertion and deletion costs of 3, length fraction of 0.8, and similarity fraction of 0.95). To estimate transcriptome contiguity and quality, contigs were used as *BLASTX* (Altschul et al. 1990) queries against the *Arabidopsis thaliana* proteins (ftp://ftp.arabidopsis.org/home/tair/Proteins/) with default settings. Coverage of the top hit reference proteins was then used as a measure of transcript completeness.

To reduce the risk that results in downstream analyses could be biased by ancient introgression between the ingroup and outgroup lineages, alignments were produced using a reference-guided approach based on two outgroup *de novo* transcriptomes, those of *S. madagascariensis* and *S. flavus*. Trimmed reads from each individual were mapped to the two outgroup reference transcriptomes separately using *CLC* with the following settings: length fraction = 0.8, similarity fraction = 0.9, automatic detection of paired end distances, mismatch cost = 2, insertion cost = 3, deletion cost = 3. BAM files for each mapping were exported and the *samtools/BCFtools* package version 1.1 (Li et al. 2009) was used for variant calling, filtering and consensus calling. BAM files were converted to pileup format using the *samtools mpileup* command, using a base quality filter of 20 and a mapping quality filter of 20. The *bcftools call* command was then used for SNP calling with a minimum read depth filter of 8. Several further filters were implemented using the *bcftools filter* command: SNPs within 3 bases of indels, with a quality below 10, and with

each allele represented by less than 2 reads for heterozygous SNPs were removed. The resulting VCF files were converted to fasta format using a custom *C++* script (provided by Bruno Nevado), indels were converted to missing data, and heterozygous SNPs were represented by IUPAC codes. Since two references, *S. madagascariensis* and *S. flavus,* were used for the mapping, two sets of alignments were produced and these were carried separately into downstream analyses.

3.2.3 Phylogenetic inference

Since incomplete lineage sorting (ILS) is likely to be high in recently diverged taxa such as the focal species, the phylogeny of the species was estimated using the multi-species coalescent-based approach of Mirarab et al. (2014), which accounts for ILS. First, for each reference-guided assembly, each single-contig alignment was used in a separate *RAxML* analysis, using the *GTRCAT* model and 100 bootstrap replicates. The bootstrap replicates and best maximum likelihood trees from these analyses were then used to produce a species tree using *ASTRAL* (Mirarab et al. 2014) with 100 bootstrap replicates using site-wise resampling. This produced a bootstrapped species topology for each of the reference-guided datasets and since ILS is likely to be high in the focal species. As a secondary approximation topology, and to determine branch lengths for the phylogeny, species tree estimation was also undertaken using a concatenation-based maximum likelihood approach. Each of the two reference-guided assemblies, *S. flavus* and *S. madagascariensis*, were concatenated separately using a custom *bash* script (available on request). Maximum likelihood tree inference was then conducted on these concatenated datasets in *RAxML* version 8 (Stamatakis 2014) using the *GTRCAT* model and 100 bootstrap replicates.

To visualise phylogenetic discordance between loci, the best maximum likelihood tree for each contig from the per-locus *RAxML* analysis was used to produce *DensiTree* plots (Bouckaert 2010). For each gene tree, nodes with bootstrap support below 75% were collapsed using the *pruneTree* function in the *phangorn* package (Schliep 2011) in *R* 3.1.2 (R Core Development Team, 2014). Gene trees with more than two nodes with less than 75% bootstrap support were removed and each tree was rooted by *S. flavus* using the *root* function in in the *APE* module (Paradis et al. 2004) in *R*. Rooted, pruned trees were then made ultrametric using the *chronos* function in the APE module (Paradis et al. 2004) in *R* with default settings. The resulting pruned, rooted and ultrametric trees were then input into *DensiTree* (Bouckaert 2010). *DensiTree* plots were produced using the consensus trees produced by *DensiTree* (in which branch lengths are averaged across all trees for a given topology) with the following settings (arced tree, consensus width = 1, consensus intensity 28.1 and default values for all other settings). *DensiTree* plots were produced in this way for contigs for each of the two reference-guided assemblies as well as for both sets of contigs combined.

To determine how much of the observed variation between gene trees was due to genuine incongruence, rather than simply lack of phylogenetic signal, Shimodaira-Hasegawa (SH) tests implemented in *CONSEL* were used on all loci from each of the two reference-guided assemblies (Shimodaira & Hasegawa 1999, 2001). The procedure first uses the phylogenetic inference program *PhyML* (Guindon & Gascuel 2003) over two runs for each contig. The first run uses an unconstrained topology, and the second constrains the topology to that of the species tree hypothesis obtained from the phylogenetic analyses (which had the same topology for both reference-guided assemblies as well as the best

maximum likelihood trees from the two concatenated reference guided assemblies – see 3.3 Results). These were run using the *GTR* substitution model, without bootstrap replicates. Site-likelihoods from these are then compared in *CONSEL* and FDR correction was applied to the *P*-values for each SH test (Benjamini & Hochberg 1995). Tests for which the FDR corrected SH-test *P*-value (*Q*-value) after FDR correction for the constrained tree is below 0.05 are considered to significantly reject the species tree. The *PhyML-CONSEL* and per-contig *RAxML* pipelines were automated using a custom *perl* and *bash* scripts.

3.2.4 Tests for introgression

To investigate possible introgression between the species as a cause of the observed incongruence (see 3.3 Results), Patterson's *D*-statistic test (Durand et al. 2011) was employed. This compares two phylogenetically incongruent site patterns of ancestral (A) and derived (B) alleles ABBA - (((A,B),B),A) and BABA - (((B,A),B),A) on a four-taxon phylogeny with the topology: (((Sp.1,Sp.2),Sp.3),Outgroup). If the incongruence is due to incomplete lineage sorting, the frequencies of these sites are expected to be equal, but in the case of introgression between Sp.3 and either Sp.1 or Sp.2, they are expected to be skewed towards the site pattern that clusters the introgressing taxa together. Block Jackknifing (with each transcript representing a single block in the context of these datasets) was then used to determine significance (Korneliussen et al. 2014). The *doAbbaBaba* function in *ANGSD* (Korneliussen et al. 2014) was used to test every phylogenetically congruent three-species subtree from the six European species using both *S. flavus* and *S. madagascariensis* as the outgroup/mapping-reference separately. This approach estimates counts of ABBA and BABA sites using base counts from Binary Alignment/Map (BAM) files applying a minimum read coverage filter of five, and a

minimum mapping quality filter of 20 to potential ABBA/BABA sites (Korneliussen et al. 2014). To minimise the effect of outgroup choice (and any bias caused by potential past introgression from the outgroup) only the tests in which both outgroups produced similar results were considered to be significant. All *P*-values were corrected for multiple testing using the False Discovery Rate method of Benjamini & Hochberg (FDR; 1995).

3.3 Results

3.3.1 Dataset

Data from *S. flavus* was assembled into 25,035 contigs and *S. madagascariensis* was assembled into 29,739 contigs over 300 bp with respective N50 values of 1,102 and 1,093 nucleotides (bp) and total lengths of 23,420,882 and 25,857,323 bp (Number of reads shown in Table 3.1). *BLASTX* searches against the *Arabidopsis* transcriptome were used to estimate transcript contiguity and transcriptome completeness, and both transcriptomes showed good coverage of the *Arabidopsis* proteins, indicating that transcript completeness was high (24.9% of *S. flavus* contigs and 24.3% of *S. madagascariensis* contigs aligned to at least 90% of their top hit; 35.92% and 36.91% of *Arabidopsis* proteins were hit by the *S. flavus* and *S. madagascariensis* assemblies respectively). Data from all eight species were mapped to the outgroup contigs and, after filtering, the *S. flavus* reference contained 22,100 transcripts and the *S. madagascariensis* reference contained 25,431 transcripts.

3.3.2 Phylogenetic inference

Using the multi-species coalescent-based method of Mirarab et al. (2014), on each of the two reference-based assemblies, as well as both of the concatenation-based analyses produced the same topology (Fig. 3.1A). To visualise the level of gene tree-species tree

incongruence across the genome, gene trees for each transcript from both of the reference-based assemblies (those used for estimation of the multi-species coalescent-based approach, above) were used to build a *DensiTree* plot (Bouckaert 2010). Gene trees are highly variable, but a sister relationship between *S. aethnensis* and *S. chrysanthemifolius* and monophyly of *S. aethnensis, S. chrysanthemifolius, S. leucanthemifolius, S. glaucus, S. gallicus* and *S. vernalis* can clearly be seen (Fig. 3.1B). The variation among gene tree topologies could be due to a lack of information in any single contig, or genuine incongruence e.g. from incomplete lineage sorting (ILS) or introgression. To determine whether data from individual contigs significantly rejected the species tree, a series of SH tests were implemented to compare the fit of the estimated species tree and an unconstrained topology to the data. This showed that 1561 (7.06%) of *S. flavus* reference-based contigs and 1918 (7.54%) of *S. madagascariensis* reference-based contigs significantly reject the species tree ($P < 0.05$) which dropped to 370 and 582 respectively after FDR correction.

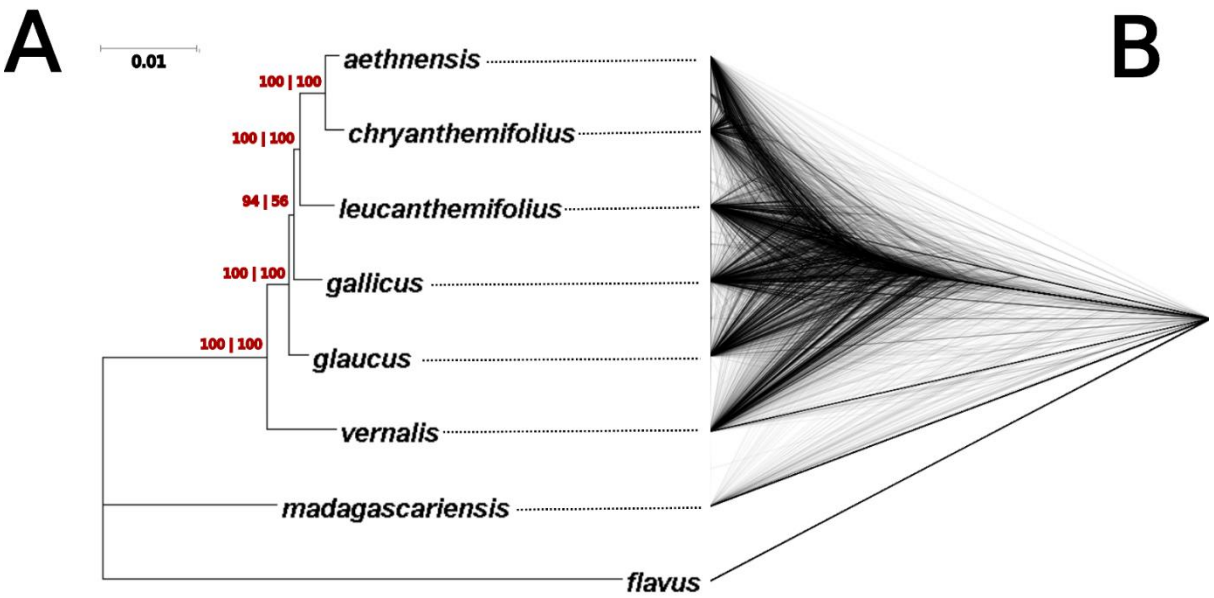

Figure 3.1: Phylogenetic reconstruction and gene tree-species tree incongruence. (A) Species phylogeny estimated using *RAxML* (Stamatakis 2014) and *ASTRAL* (Mirarab et al. 2014), both of which produced the same topology. Branch lengths are the mean of those produced using *RAxML* on both reference-based datasets. For each node, bootstrap support from the method of Mirarab et al. (2014) for the *S. flavus* and

***S. madagascariensis* reference-based assemblies are shown before and after the bar respectively. (B) A *DensiTree* plot of maximum likelihood gene trees for all contigs. For each gene tree, nodes with less than 75% bootstrap support are collapsed and gene trees which subsequently contain more than two polytomies are excluded. For each unique topology in amongst gene trees, branch lengths are averaged amongst all gene trees showing that topology. Shown are results using trees from both reference-based assemblies combined.**

3.3.3 Detection of introgression

The high level of gene tree-species tree incongruence could be due to either incomplete lineage sorting (ILS) or introgression. To differentiate between these possibilities a D-statistic approach was used, in which each species-tree-congruent triplet of ingroup species was used in an ABBA-BABA test (Durand et al. 2011). For each triplet of ingroup species, this was performed once with each of the outgroups *S. madagascariensis* and *S. flavus* as mapping reference and outgroup used to polarise SNPs. The results indicated a large number of introgression events amongst the Mediterranean species had occurred, which were widely distributed across the tree (Table 3.2). Assuming the species tree topology is correct, introgression was inferred in 12/20 tests using *S. flavus* as a reference, and 11/20 tests using *S. madagascariensis*. There was a high level of agreement between the results using each reference, with 10/20 being significant with both references, and 7/20 being non-significant in both, thus only 3 tests showed disagreement between the references (Table 3.2). There was also a strong and highly significant correlation between the *D*-statistic estimates from the two references (Pearson's product moment correlation test: $P = 9.4 \times 10^{-8}$, $R = 0.895$; Figure 3.2).

Table 3.2: ABBA-BABA test results. Results are shown for each for all phylogenetically congruent triplets of species using both outgroups as the mapping reference and outgroup. Species abbreviations are as follows: aeth: *S. aethnensis*, chry: *S. chrysanthemifolius,* leuc: *S. leucanthemifolius*, gall: *S. gallicus*, glau: *S. glaucus*, vern: *S. vernalis*. Species inferred to have introgressed in tests using both outgroups are highlighted in bold.

Test species			Mapping reference/outgroup			
			S. flavus		*S. madagascariensis*	
Sp. 1	Sp. 2	Sp. 3	*D* ± SE	*Q*-value	*D* ± SE	*Q*-value
chry	aeth	vern	-0.01 ± 0.007	0.208	0.003 ± 0.007	0.775
gall	aeth	**vern**	-0.027 ± 0.006	<0.001	-0.053 ± 0.006	<0.001
chry	aeth	gall	-0.001 ± 0.006	0.937	0.001 ± 0.006	0.937
gall	chry	**vern**	-0.014 ± 0.006	0.036	-0.053 ± 0.006	<0.001
leuc	aeth	gall	0.01 ± 0.006	0.094	0.007 ± 0.005	0.295
leuc	chry	gall	0.017 ± 0.006	0.004	0.006 ± 0.006	0.388
gall	leuc	vern	-0.004 ± 0.006	0.547	-0.005 ± 0.006	0.498
leuc	aeth	**vern**	-0.02 ± 0.006	0.002	-0.035 ± 0.006	<0.001
chry	aeth	**leuc**	-0.038 ± 0.006	<0.001	-0.027 ± 0.006	<0.001
leuc	chry	vern	-0.011 ± 0.006	0.094	-0.044 ± 0.006	<0.001
glau	aeth	**vern**	-0.044 ± 0.006	<0.001	-0.054 ± 0.006	<0.001
chry	aeth	glau	-0.013 ± 0.006	0.067	-0.005 ± 0.007	0.528
glau	chry	**vern**	-0.035 ± 0.006	<0.001	-0.065 ± 0.006	<0.001
gall	aeth	**glau**	-0.075 ± 0.005	<0.001	-0.093 ± 0.006	<0.001
gall	chry	**glau**	-0.071 ± 0.006	<0.001	-0.096 ± 0.005	<0.001
gall	leuc	**glau**	-0.074 ± 0.005	<0.001	-0.088 ± 0.005	<0.001
glau	gall	vern	-0.019 ± 0.006	0.002	-0.01 ± 0.006	0.129
leuc	aeth	glau	0 ± 0.006	0.941	-0.004 ± 0.005	0.523
leuc	chry	glau	0.009 ± 0.006	0.129	-0.01 ± 0.006	0.103
glau	leuc	**vern**	-0.021 ± 0.006	<0.001	-0.017 ± 0.006	0.007

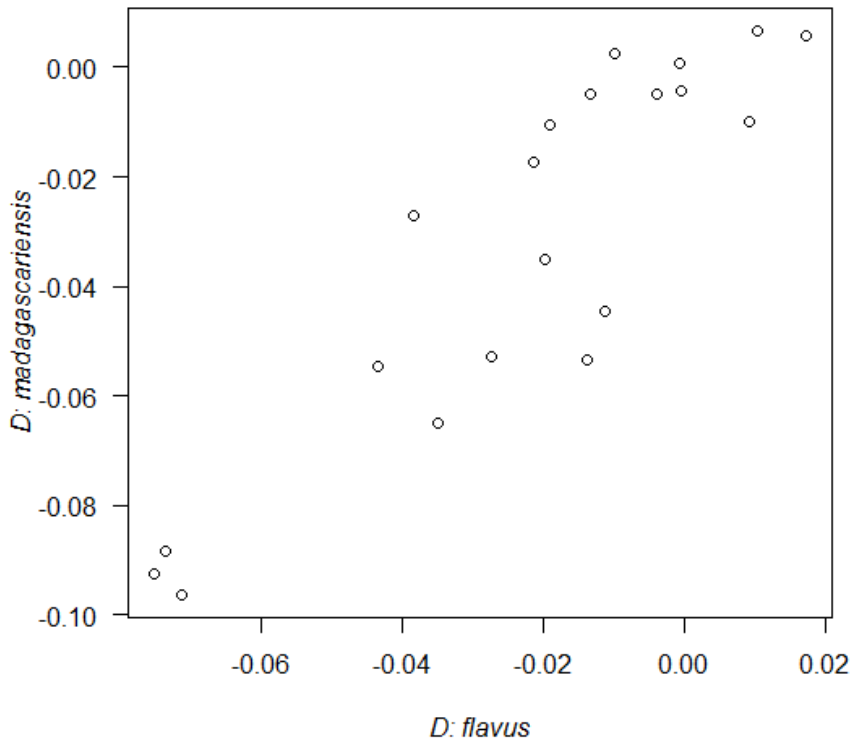

Figure 3.2: Relationship between *D*-statistic estimates for ABBA-BABA tests using *S. flavus* and *S. madagascariensis* as the outgroup and mapping reference.

3.4 Discussion

3.4.1 Species level phylogeny and gene tree-species tree incongruence

Despite the large amount of work on the Mediterranean *Senecio* species complex (e.g. Brennan et al. 2009; Chapman & Abbott 2010; Hegarty et al. 2006; Kim et al. 2008; Lowe & Abbott 2004; Pelser et al. 2012) a fully resolved phylogenetic history of the species had previously remained elusive (Comes & Abbott, 2001; Pelser et al. 2007). This is mirrored by long-standing difficulties experienced by taxonomists in species identification and establishing satisfactory species delimitations amongst the Mediterranean species complex (Alexander 1979; Crisp 1972). The results in this chapter may go some way to explaining these difficulties, since a proportion of genes significantly rejected the species

tree. This incongruence could clearly be seen when individual gene trees were combined into a *DensiTree* plot (Fig. 3.1B). Since the species are recently diverged, and thus incomplete lineage sorting (ILS) is likely to be high in the species, using a method such as that of Mirarab et al. (2014) which takes ILS into account, is likely to produce a more accurate species topology than most methods. Nevertheless, the method does not account for gene flow between the species. Interspecific gene flow can create false monophyletic relationships, as well as making inference of correct monophyletic relationships more likely when gene flow is between sister species (Leaché et al. 2014) and evidence of extensive gene flow has been found by the analyses in this thesis. Therefore, the phylogenetic hypotheses presented here, as well as future phylogenetic efforts in these species using similar methods, should be taken with some level of caution.

Those caveats notwithstanding, the phylogeny inferred here has important implications. Perhaps the most notable feature was the fact that *S. aethnensis* and *S. chrysanthemifolius* were sister species (with respect to the species samples in this study). This is important because *S. aethnensis* and *S. chrysanthemifolius* have been previously postulated to be a case of recent ecological speciation (Osborne et al. 2013). Both species have very limited geographical ranges which abut in an altitude-associated hybrid zone where they hybridise extensively. Nevertheless, they are highly phenotypically distinct and there is evidence for divergent selection between them and low levels of both pre- and post-zygotic reproductive isolation (Ross 2010; Ross et al. 2012; Brennan et al. 2009, 2014). Thus their apparent monophyly, at least relative to the other species sampled in this study, supports the possibility that they may have speciated *in situ* as a result of their differential adaptation to high and low altitude habitats. The sister relationship should be taken with

some amount of caution however. Firstly, since extensive enough introgression can cause the incorrect inference of sister relationships between species, and *S. aethnensis* and *S. chrysanthemifolius* are known to undergo introgressive hybridisation, then it is possible that this sister relationship could be incorrect (Leaché et al. 2014). This scenario assumes such a high level of gene flow that genetic swamping of one species with the other has occurred (Kutschera et al. 2014; Leaché et al. 2014). However, since the phylogenetic support for this relationship is so strong, then what remains of the original recipient taxon today if this was the case represents a very small proportion of the genome. Thus the species as they exist today, are essentially sister species across the vast majority of the genome with respect to the other species analysed. A more important point regarding the *S. aethnensis* – *S. chrysanthemifolius* sister relationship is that data was not available for all species in the clade. Previous analyses based on plastid DNA and allozymes have found that *S. rupestris*, a species found in mountainous regions of central and southern Europe clusters with *S. aethnensis* in phylogenies, although there was very low statistical support for this relationship (bootstrap support < 50%; Abbott et al. 2002). Thus, a focus of future work should be to produce a high-throughput phylogenetic analysis of the clade including all species in the Mediterranean *Senecio* species complex, particularly *S. rupestris*, to confirm or deny the sister relationship between *S. aethnensis* and *S. chrysanthemifolius.*

It is also worth noting that the only node with bootstrap support below 100%, was that partitioning *S. glaucus* from the clade containing *S. gallicus, S. leucanthemifolius, S. chrysanthemifolius* and *S. aethnensis.* The phylogenetic positions of *S. glaucus, S. gallicus* and *S. leucanthemifolius* are also the most difficult to discern when viewing the *DensiTree* plot. These three species are all widespread species with partially overlapping ranges. It is

possible that more frequent episodes of hybridisation between the more widely distributed species in the clade, could have extensively muddied the phylogenetic waters in Mediterranean *Senecio*. Indeed, there is strong evidence for introgression between *S. gallicus* and *S. glaucus* from the ABBA-BABA tests.

3.4.2 Introgression is widespread in the group

Gene tree-species tree incongruence can have many sources, which can be broadly divided into coalescent processes: the incomplete sorting of ancestral variation; and reticulate processes: which include introgressive hybridisation, hybrid speciation and vector-mediated horizontal gene transfer. The results in this chapter provide evidence that at least part of the explanation for the gene tree-species tree incongruence identified is extensive introgressive hybridisation in the clade.

The system of multiple ABBA-BABA tests used here can provide some insight with respect to the phylogenetic position of introgression events, although the exact phylogenetic position and timing of introgression can often not be inferred. For example, assuming the species-tree topology is correct, multiple tests support introgression between *S. vernalis* and each of *S. leucanthemifolius, S. gallicus* and *S. glaucus*. This result could be interpreted in several ways. Firstly, it could represent separate episodes of introgression between *S. vernalis* and each of these species. Secondly, it could result from a more ancient introgression event between the ancestor of *S. vernalis* and the common ancestor of *S. leucanthemifolius, S. gallicus, S. glaucus, S. aethnensis* and *S. chrysanthemifolius* with introgressed material subsequently being lost in *S. aethnensis* and *S. chrysanthemifolius*. And thirdly, it could be due to introgression between *S. vernalis* and only one of the three species: the remaining significant tests in this case resulting from either subsequent

introgression between the recipient species and the other two, or to introgression into *S. vernalis* of genetic polymorphisms shared by *S. leucanthemifolius, S. gallicus* and *S. glaucus*. Furthermore, these tests do not preclude introgression between *S. vernalis* and *S. aethnensis/chrysanthemifolius*. This could still have occurred, but if so, it occurred to a greater extent between *S. vernalis* and the other three species in the clade. A final caution regarding the interpretation of ABBA-BABA tests is that ancestral population structure could potentially provide false-positive results. It has been shown that some very specific cases of population structure can give rise to *D*-statistic patterns which are indistinguishable from introgression (see Durand et al. 2011; Eriksson & Manica 2012) although this seems an unlikely source of such a large number of significant tests.

Overall, while the interpretation of multiple ABBA-BABA tests can be ambiguous in terms of the exact phylogenetic position of introgression it is clear that introgression has been widespread in the clade. Thus, while the individual ABBA-BABA tests are dependent on the correct phylogeny being known, the overarching conclusion of widespread introgression in the clade is robust to phylogenetic uncertainty.

In a genus with such high levels of interspecific hybridisation, finding an outgroup with no previous contact with the focal species is challenging. The lineages containing both *S. madagascariensis* and *S. flavus* may have experienced historical hybridisation with the clade containing the focal species (Kadereit et al. 2006; Pelser et al. 2012). This is potentially problematic because the *D*-statistic approach used could be affected by unknown introgression from the outgroup used to polarise genetic variation and estimate locus-specific evolutionary rates, because this would introduce uncertainty of the ancestral state (Durand et al. 2011). For this reason, two species which were not closely related to

each other (Pelser et al. 2007) were used for the *de novo* reference transcriptomes and outgroups for these tests. The high level of agreement between the outgroup replicates is encouraging: 17 out of 20 ABBA-BABA tests were either significant with both outgroups, or neither. Only tests which were significant in both after multiple test correction were considered as evidence for introgression, thus the results are expected to be reliable. Similarly, there were very high levels of correlation between the results using each outgroup in the *D*-statistic values provided by the ABBA-BABA tests. The genus *Senecio* has experienced a large number of known hybridisation events (Comes & Abbott 1999; Kadereit et al. 2006; Pelser et al. 2012) and it is possible that this is common throughout the genus, so it would be challenging to find outgroups for which no introgression since the split with the focal species could be certain. These results underline the fact that such problems can be ameliorated, and a higher level of certainty can be reached, by using multiple outgroups.

One of the specific conclusions from the *D*-statistic analysis matches those reached in previous work. A secondary contact zone has been reported between *S. glaucus* and *S. vernalis* in Israel (Comes and Abbott, 1999) in which introgression has been inferred from sharing of cpDNA haplotypes between the species despite an ITS phylogeny placing them in distinct well supported clades. The *S. glaucus* and *S. vernalis* accessions used in this chapter were from geographically distant populations (Morocco and Cyprus respectively; Table 3.1) suggesting that sharing of introgressed genetic material between the species is not restricted to contemporary parapatric populations in the Near East. The most parsimonious interpretation (i.e. the scenario which requires the fewest number of episodes of introgression) is reported in Fig. 3.3 and Table 3.3. This does not imply,

however, that this is necessarily the most likely scenario. Indeed, it is quite possible that the history of introgression in the clade is far more complex than this, and involves multiple episodes of introgression between each lineage or consistently low-level introgression throughout their evolution.

3.4.3 Conclusions

Overall, it can be concluded that, despite their phenotypic differences, probable local adaptation, and habitat preference differences, the clade as a whole has experienced widespread gene flow throughout a substantial portion of its evolutionary history. Indeed, every species examined was found to have introgressed with at least one other species when the results of this chapter and Chapter 2 are considered. What largely remains to be seen is the evolutionary role it plays in the species. Unfortunately, this dataset is unsuited to identifying the specific loci which have introgressed between the species since the *D*-statistic is likely to be dominated by stochastic variation in the short regions of sequence data produced by RNA-seq (Martin et al. 2015). However, the completion of the *Senecio* Genome Project (T. Batstone, M.A. Chapman, O.G. Osborne, D.A. Filatov, R.J. Abbott, and S.J. Hiscock, in preparation) will give access to longer genomic windows, which could be used for this purpose with additional resequencing of the species investigated here. This would allow questions regarding the role of gene flow to be addressed. For example, by comparing signatures of selection in introgressed versus non-introgressed loci, it could be determined whether introgression is ever adaptive in these species, or if the introgressed regions are likely to be those which are functionally similar and selectively neutral between the species, such that their introgression has no adaptive effect. The complex interactions of gene flow and selection; which may all vary spatially, temporally and between loci; can

combine to lead to diverse outcomes at the species level. To what extent species can adapt to their environments, whether speciation will or will not occur, and to what extent species boundaries are maintained after initial divergence depends on these factors, and understanding their interaction during species diversification represents one of the most important challenges in evolutionary biology (Nosil et al. 2009; Seehausen 2004). The results in this chapter, finding as they do far more widespread introgression than was previously known in this clade, are an important step towards establishing the Mediterranean *Senecio* complex as one of the foremost systems in which to study the basis and evolutionary consequences of gene flow during species diversification.

Table 3.3: Possible interpretations of results from introgression analyses

Hypothesis	Most parsimonious interpretation	Evidence from this study	Alternative interpretations	Evidence from previous studies
A	Introgression between the *S. aethnensis* and *S. chrysanthemifolius* lineages	Significant LRT for introgression between *S. aethnensis* and *S. chrysanthemifolius* – Chapter 2	n/a	Well documented in the literature. See e.g. Brennan et al., 2009
B	Introgression between the *S. leucanthemifolius* and *S. chrysanthemifolius* lineages. Including introgression of variation shared by *S. aethnensis and S. chrysanthemifolius* into *S. leucanthemifolius.*	An excess of shared incongruent SNPs between *S. leucanthemifolius* and *S. chrysanthemifolius* relative to *S. aethnensis.*	n/a	Not previously reported
C	Introgression between the *S. gallicus* and *S. glaucus.*	An excess of shared incongruent SNPs between *S. gallicus* and *S. glaucus* relative to *S. leucanthemifolius, S. chrysanthemifolius* and *S. aethnensis.*	Incorrect tree topology, since the node uniting *S. gallicus, S. leucanthemifolius, S. chrysanthemifolius* and *S. aethnensis* had low bootstrap support.	Not previously reported
D	Introgression between the *S. vernalis* lineage and the common ancestor of all other species in the clade, subsequent loss of introgressed variation in *S. aethnensis* and *S. chrysanthemifolius.*	An excess of shared incongruent SNPs between *S. vernalis* and *S. gallicus* and *S. glaucus* relative to *S. aethnensis* and *S. chrysanthemifolius.* An excess of shared incongruent SNPs between *S. vernalis* and *S. glaucus* relative to *S. leucanthemifolius.* An excess of shared incongruent SNPs between *S. vernalis* and *S. leucanthemifolius* relative to *S. aethnensis.*	Separate introgression between *S. vernalis* and *S. leucanthemifolius, S. gallicus* and *S. glaucus.* Introgression of *S. vernalis* alleles into *one of the species* and subsequent introgression of that material onto the others. While the scenario presented in the "most parsimonious interpretation" column requires the fewest episodes of introgression, the requirement that the introgressed material is subsequently lost in *S. aethnensis* and *S. chrysanthemifolius* makes this scenario seem somewhat less plausible than independent introgression events between *S. vernalis* and *S. leucanthemifolius, S. glaucus* and *S. gallicus.*	Evidence of introgression between *S. vernalis* and *S. glaucus* (Comes & Abbott, 1999).

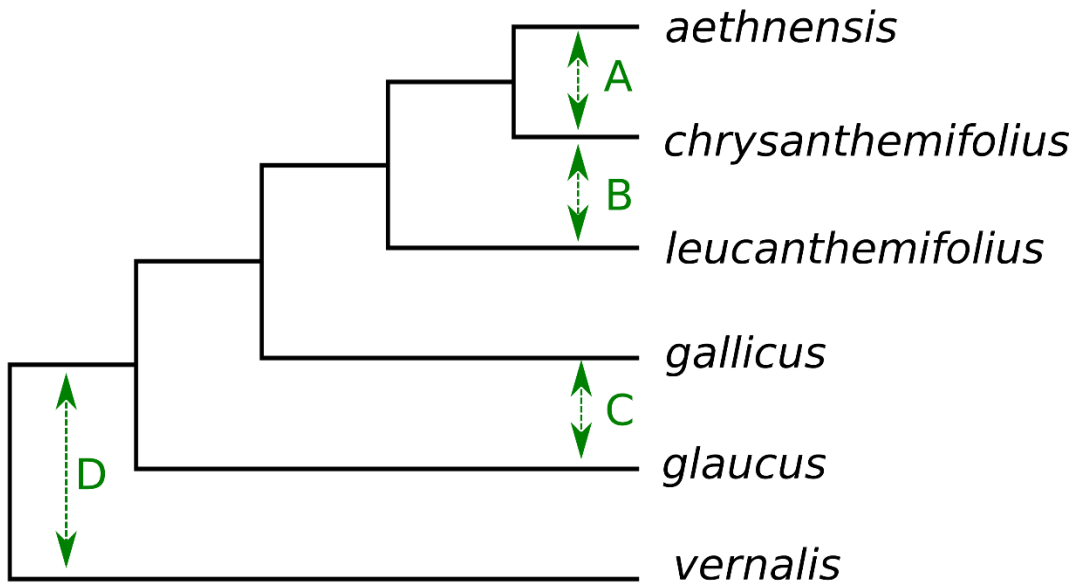

Figure 3.3: The scenario explaining the results of introgression analyses in this chapter as well Chapter 2 which requires the fewest number of episodes of introgression. Green arrows represent introgression events and letters refer to table 3.3. Branch lengths are arbitrary. See Table 3.3 for justification.

Chapter 4: The rate of positive selection in *Senecio* relative to other plant taxa.

Preface

This chapter was in preparation for submission as a research article at the time of printing. All analytical work and writing is my own with following exceptions. Some analytical scripts were provided by Bruno Nevado (noted in the text). *Silene* RNA extractions were performed by Dmitry Filatov.

4.1 Introduction

The genus *Senecio* is one of the largest plant genera, and is highly morphologically and ecologically diverse (Pelser et al. 2011). Even within the group of very closely related species investigated in Chapter 3, there is a vast level of diversity in habitat preference. For example *S. glaucus* occurs in desert-like environments (Comes & Abbott 1999), whereas *S. aethnensis* is found in extreme alpine habitats near the summit of Mount Etna (Chapman et al. 2013). However, despite their extensive divergence they have experienced substantial gene flow, as shown in Chapter 3. Thus, for this phenotypic divergence to evolve, especially in the face of gene flow, it might be expected that *Senecio* has experienced a high rate of diversifying selection. However, this hypothesis has not been tested, and few candidate genes under selection have been identified.

Codon-based phylogenetic methods to detect selection provide a powerful likelihood framework in which the targets of positive selection, and the strength of selection can be assessed (Nielsen & Yang 1998; Yang et al. 2000; Yang & Nielsen 2002). However, the small number of species analysed using them in Chapter 2 lack the power for a robust test of positive selection on a gene-by-gene basis. Of the few other studies which have investigated selection in *Senecio* (Brennan et al. 2009; Ross et al. 2012; Chapman et al. 2013; Muir et al. 2013; Oberprieler et al. 2015; Roda et al. 2013), all have studied just one or two species and the majority have focussed on the species-pair of *S. aethnensis* and *S. chrysanthemifolius*. Of these, the only studies which have attempted to identify the loci under selection, rather than using anonymous markers or ecological methods, have both focussed on *S. aethnensis* and *S. chrysanthemifolius* (Chapman et al. 2013; Muir et al. 2013). Perhaps more significantly, no studies have made any quantitative comparison of rates of positive selection in *Senecio* to

that in other genera. To make any inference about levels of selection being high, low or average in any taxon, it is essential to apply the same methods of inference to comparative datasets from other taxa, ideally on the same set of genes.

Codon-based tests of selection operate on the principle that if synonymous sites are assumed to evolve neutrally, the ratio of non-synonymous to synonymous divergence, d*N*/d*S*, can be used to infer the direction and strength of selection, with an increase in non-synonymous divergence implying positive (directional) selection, and a deficit of non-synonymous divergence implying purifying (negative, or stabilising) selection. However, since most codons within a gene are expected to be under purifying selection, the d*N*/d*S* ratio of a gene is almost never positive, even if some codons are under positive selection. For this reason, models have been developed which allow d*N*/d*S* to vary among codons within a gene (Nielsen & Yang 1998; Yang et al. 2000; Yang & Nielsen 2002; Yang 2007). This allows statistical comparison of models which allow d*N*/d*S* above one at some sites with models which do not, providing a powerful test of positive selection.

This chapter uses these methods to produce a genome-wide analysis of selection across eight species of *Senecio.* To determine a "base-line" to compare the level of selection in *Senecio* to, similar datasets from eight other genera were analysed by the same methods. These analyses aimed to answer the following questions. i) Which genes have been under positive selection during the diversification of the *Senecio* species analysed? ii) What proportion of genes has been under positive selection in *Senecio*? iii) How does the proportion of genes under selection in *Senecio* compare to proportions in plants in general? iv) Are genes in *Senecio* under similar selective regimes as their orthologues in other plant taxa?

4.2 Materials and methods

4.2.1 Data collection

Transcriptome data was used from *Senecio* and species from eight other plant genera, some of which were published datasets and some of which were sequenced *de novo.* For *Senecio* species the raw data was that used in Chapter 3 (see Chapter 3 methods for details). For one comparison genus, *Silene*, a *de novo* dataset was sequenced. A single mature leaf and flower bud from each species was powdered in liquid nitrogen with a pestle and mortar and the *Qiagen* (Qiagen, Crawley, UK) *Plant RNeasy* kit was used for extraction following the manufacturer's guidelines including the optional DNase digestion steps to remove DNA contamination. For all species, paired-end libraries were constructed. These were multiplexed and sequenced using the Illumina Hi-seq sequencing platform with read lengths of 100 bp. Other comparison genera were obtained from the Short Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra). The SRA was searched for plant genera with Illumina RNA-seq data for at least 6 species. Species which were known to be of hybrid origin were not used since they would violate the phylogenetic model used in the analyses of selection. Genera which contained data for six or more species were carried into downstream analyses (Table 4.1). These were *Flaveria* (8 spp.)*, Glycine* (6 spp.)*, Helianthus* (7 spp.)*, Linum* (8 spp.)*, Oryza* (8 spp)*, Populus* (6 spp.) and *Solanum* (8 spp.)

Table 4.1: Sources of data used in this chapter. [1] Short Read Archive. PE = paired-end. SE = single end

Genus	Species	Sequenced *de novo* or from databases	Raw read length	Paired or single end	Number of raw reads	Number of reads after trimming and subsetting	SRA[1] accession number
Flaveria	*anomala*	database	100	SE	51,845,372	35,428,344	SRR1165196
Flaveria	*bidentis*	database	90	PE	23,089,000	22,962,970	SRR1141027
Flaveria	*brownii*	database	90	PE	21,457,762	21,388,314	SRR1141029
Flaveria	*chlorifolia*	database	100	SE	83,862,797	58,225,242	SRR1165206
Flaveria	*pubescens*	database	90	PE	20,703,102	20,628,618	SRR1141037
Flaveria	*ramosissima*	database	100	SE	61,878,200	47,418,685	SRR1166363
Flaveria	*robusta*	database	100	SE	62,212,599	44,479,238	SRR1166375
Flaveria	*trinervia*	database	90	PE	27,345,748	27,195,006	SRR1141040
Glycine	*canescens*	database	86	SE	7,018,982	6,872,257	SRR1171850
Glycine	*clandestina*	database	86	SE	8,121,993	8,015,758	SRR1171830
Glycine	*max*	database	100	PE	72,891,652	71,494,510	SRR1174226
Glycine	*soja*	database	76	PE	59,187,378	58,089,068	SRR1211016
Glycine	*syndetika*	database	100	SE	14,361,584	13,168,561	SRR452306
Glycine	*tomentella*	database	97	SE	14,857,517	13,741,266	SRR452313
Helianthus	*annuus*	database	76	PE	62,929,328	54,799,892	SRR1685780
Helianthus	*argophyllus*	database	100	PE	44,082,252	38,228,832	SRR826582
Helianthus	*bolanderi*	database	100	PE	128,877,846	50,000,000	SRR1043153
Helianthus	*exilis*	database	100	PE	46,318,096	46,083,702	SRR1041637
Helianthus	*petiolaris*	database	100	PE	55,830,278	41,871,220	SRR826626
Helianthus	*praecox*	database	100	PE	67,371,166	49,370,278	SRR826631
Helianthus	*debilis*	database	100	PE	52,462,890	41,908,184	SRR826604
Linum	*bienne*	database	90	PE	22,761,090	22,515,366	SRR957663

Table 4.1 cont.

Linum	*grandiflorum*	database	90	PE	18,920,884	18,828,614	SRR957662
Linum	*hirsutum*	database	90	PE	29,650,598	29,191,902	SRR957659
Linum	*lewisii*	database	90	PE	29,195,398	28,751,796	SRR957658
Linum	*macraei*	database	90	PE	26,639,782	26,466,554	SRR957660
Linum	*perenne*	database	90	PE	29,971,822	29,778,462	SRR957661
Linum	*tenuifolium*	database	90	PE	27,939,962	27,486,908	SRR957665
Linum	*usitatissimum*	database	90	PE	26,397,156	26,210,298	SRR957669
Oryza	*barthii*	database	100	PE	486,397,404	50,000,000	SRR1170762
Oryza	*glumipatula*	database	100	PE	467,047,814	50,000,000	SRR1174772
Oryza	*granulata*	database	100	PE	505,971,512	50,000,000	SRR1178922
Oryza	*meyeriana*	database	90	PE	162,133,290	50,000,000	SRR1685731
Oryza	*nivara*	database	75	PE	500,419,328	50,000,000	SRR1171631
Oryza	*officinalis*	database	100	PE	429,565,428	50,000,000	SRR1179195
Oryza	*rufipogon*	database	90	PE	418,194,358	50,000,000	SRR1220645
Oryza	*sativa*	database	150	PE	123,387,512	50,000,000	SRR1592601
Populus	*euphratica*	database	90	PE	132,575,472	50,000,000	SRR901769
Populus	*pruinosa*	database	100	PE	50,025,426	38,853,992	SRR1046128
Populus	*tomentosa*	database	100	PE	119,716,602	50,000,000	SRR1508229
Populus	*tremula*	database	100	PE	68,548,278	62,219,480	ERR260313
Populus	*tremuloides*	database	76	SE	6,667,839	6,504,011	SRR540223
Populus	*trichocarpa*	database	76	PE	52,687,394	43,416,772	SRR1660793
Senecio	*aethnensis*	*de novo*	100	PE	15,181,318	14,809,352	SRA096168
Senecio	*chrysanthemifolius*	*de novo*	100	PE	12,611,162	12,325,626	SRA096168
Senecio	*madagascariensis*	*de novo*	100	PE	34,606,224	33,417,248	SRP069830
Senecio	*flavus*	*de novo*	100	PE	32,273,892	31,606,490	SRP069830

Table 4.1 cont.

Senecio	*gallicus*	*de novo*	100	PE	19,108,048	17,985,830	SRP069830
Senecio	*glaucus*	*de novo*	100	PE	31,246,290	30,440,732	SRP069830
Senecio	*leucanthemifolius*	*de novo*	100	PE	33,166,514	32,356,168	SRP069830
Senecio	*vernalis*	*de novo*	100	PE	36,444,014	35,483,252	SRA096168
Silene	*acaulis*	*de novo*	100	PE	21,668,592	20,121,522	N/A
Silene	*diclinis*	*de novo*	100	PE	27,974,916	26,963,318	N/A
Silene	*dioica*	*de novo*	100	PE	15,804,142	15,006,734	N/A
Silene	*gallica*	*de novo*	100	PE	21,349,926	19,953,156	N/A
Silene	*vulgaris*	*de novo*	125	PE	39,393,414	39,052,176	N/A
Silene	*latifolia*	*de novo*	100	PE	12,920,504	12,206,788	N/A
Solanum	*berthaultii*	database	150	SE	14,145,325	13,639,000	ERR185927
Solanum	*habrochaites*	database	75	PE	70,074,204	68,964,680	SRR521349
Solanum	*incanum*	database	100	SE	50,735,884	50,398,819	SRR1087902
Solanum	*melongena*	database	100	PE	179,527,276	50,000,000	SRR1291243
Solanum	*pennellii*	database	81	PE	6,728,944	6,573,616	SRR786556
Solanum	*phureja*	database	104	PE	36,864,692	32,135,928	SRR122109
Solanum	*torvum*	database	75	PE	54,884,090	42,434,626	SRR1104128
Solanum	*dulcamara*	database	100	SE	41,251,639	40,805,147	SRR799447

4.2.2 Dataset preparation

To ensure all Illumina adaptors were removed from raw reads, *CUTADAPT* version 1.4 (Martin 2011) was used. Raw reads were then quality trimmed using *Trimmomatic* version 0.32 (Bolger et al. 2014). Coverage was highly variable between SRA datasets, so for species with over 100 million reads, 50 million were randomly subsampled using a custom *Perl* script.

For each genus, the species with the most paired-end reads remaining after trimming and filtering were selected for *de novo* assembly of a genus reference transcriptome. Reads were assembled using *TRINITY* version r20140413pl using a minimum transcript length of 300 bp and default settings for all other options. Reads from all congeneric species were then aligned to this genus-specific *de novo* reference transcriptome using *BWA* version 0.7 (Li & Durbin 2009) using the *bwa-mem* algorithm. Sequence Alignment Format (SAM) mapped reads were converted to Binary Alignment Format (BAM) using *samtools* version 1.1. *GATK* version 3.1 (DePristo et al. 2011) was used to realign mapped reads around indels.

SNP calling was then performed using the *samtools/bcftools* pipeline version 1.1. For each species individually, BAM files were converted to pileup format using the *samtools mpileup* command, using a base quality filter of 20 and a mapping quality filter of 20. The *bcftools call* command was then used for SNP calling with a minimum read depth filter of 8. Several further filters were implemented using the *bcftools filter* command: SNPs within 3 bases of indels, with a variant quality below 10, and heterozygous SNPs with alleles represented by less than 2 reads were removed. The resulting VCF files were converted to fasta format using a custom *C++* script (provided by Bruno Nevado), indels and regions which failed the

depth filter were converted to missing data, and heterozygous SNPs were represented by IUPAC codes. For each transcript alignment, species with more than 50% missing data were removed. Heterozygosity for each transcript of each species were estimated from the number of heterozygous sites per non-missing-data site.

Coding region annotation was performed in *TRANSDECODER* (Haas et al., 2013) with default settings, an approach which identifies open reading frames *de novo*. To reduce the chance of misaligned paralogues and chimeric transcripts in the dataset, only genes for which a single transcript was identified by *TRINITY* and a single best ORF was identified by *TRANSDECODER* were used in subsequent analyses. These datasets were further split into coding sequence datasets, containing only coding regions of each transcript (hereafter referred to as CDS datasets) and a dataset containing only 4-fold degenerate and non-coding regions (hereafter referred to as the neutral datasets). For the CDS datasets, codons which contained missing data in 50% of species were removed.

To determine the identity of the genes which each transcript encoded, a Conditional Reciprocal Best BLAST (*CRB-BLAST*; Aubry et al. 2014) approach was taken. Annotating genes from different species using a single reference and identical settings could be affected by distance from each species to the reference. This method ameliorates this problem by using a flexible sequence similarity cut-off. *BLASTX* searches (Altschul et al. 1990) are conducted between each assembled *de novo* reference transcriptome and the *Arabidopsis thaliana* proteome (TAIR10; ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/TAIR10_pep_20101214) and the *E*-value cut-off for each genus is adjusted to take into account its relatedness to *Arabidopsis* (Aubry et al. 2014). The *BLAST* searches are conducted in both directions (i.e.

with the reference sequences as the query and *Arabidopsis* as the database and vice versa). *De novo* transcripts were then assigned the identity of an *Arabidopsis* protein when both the transcript and the protein are each other's top *BLAST* hit.

4.2.3 Phylogenetic analysis

To determine the phylogeny of each taxon, maximum likelihood based phylogeny inference was performed on each concatenated neutral dataset using *RAxML* version 8 (Stamatakis 2014). The *GTR+Gamma* model was used, with 100 bootstrap replicates and a thorough maximum likelihood search. The best maximum likelihood tree found from this analysis is henceforth referred to as the species tree for each genus. Gene trees can differ from species trees due to introgression, ILS or alignment errors. Since the phylogeny is assumed to be known for subsequent analyses of selection, and incorrect phylogenies can marginally bias their results (Pie 2006), genes for which the gene tree was preferred over the species tree were then identified and excluded from downstream analyses. This was achieved using the method of Shimodaira & Hasegawa (1999). For all genes in each CDS dataset, phylogenetic inference was performed using *PhyML* version 3.1 (Guindon & Gascuel 2003) using the *GTR* model. This was performed in two runs, once constraining the topology to the species tree, and once with no topological constraint. Per-site likelihoods for the two trees produced in this analysis were then compared in a Shimodaira-Hasegawa (SH) test (Shimodaira & Hasegawa 1999) implemented in *CONSEL* version 1.2 (Shimodaira & Hasegawa 2001). *P*-values for all SH tests were then corrected for multiple testing using the False Discovery Rate method (FDR; Benjamini & Hochberg 1995).

4.2.4 Detection of positive selection

d*N*/d*S* based phylogenetic tests of selection can be biased by several factors, ranging from sequencing or bioinformatic error, to violations of the phylogenetic model. They are also ineffective with insufficient data. Therefore prior to this analysis, several filters were used to remove uninformative genes, or those likely to provide incorrect inferences of selection from the CDS-datasets. Firstly, genes for which sequence data was available for less than 5 species and those shorter than 33 codons were removed, as were alignments which contained no variation.

To reduce the chance of false positive inferences of selection due to erroneous SNPs caused by incorrect read mapping, three approaches were taken. Genes with extreme values of summed synonymous divergence (d*S*) are likely to represent incorrect mapping in one or more of the individuals in the alignment, so outlier genes with respect to d*S* (estimated using the *M0* model in *PAML;* Yang 2007) for each genus were removed from the dataset. Outlier genes were identified as those with d*S* greater than $Q3 + 1.5 \times IQR$, where *Q3* is the third quartile for d*S* and *IQR* is the interquartile range. Genes with extreme values for heterozygosity may represent the incorrect mapping of reads from more than one genomic locus to a single *de novo* transcript. Therefore, genes which were an outlier for heterozygosity in any species were removed. As with d*S*, outliers were defined as those with heterozygosity greater than $Q3 + 1.5 \times IQR$. Gene duplicates whose origin predates the divergence of the species, but which are incorrectly mapped to a single *de novo* transcript, would be expected to show an incongruent signal. For this reason, the results of the SH tests for disagreement with the species topology were used to remove genes for which there was less than 5% probability after FDR of the species tree being correct. This

filter also serves a second purpose, since the models used in the phylogenetic tests of selection assume a specified phylogenetic model (i.e. the species tree), genes which have been transmitted non-vertically, for example through introgressive hybridisation or vector-mediated horizontal gene transfer, also violate the test, and are likely to show the same phylogenetically incongruent signal detected by the SH test.

To identify genes under positive selection in each genus, the d*N*/d*S* based approach implemented in *PAML* version 4.8 (Yang 2007) was applied to every transcript in the reduced CDS-datasets following filtering. Two models were used. The *M7* model, which models codon-wise selection pressure as a beta distribution with d*N*/d*S* varying from 0 to 1, and the *M8* model, which includes an additional class for positively selected codons, with d*N*/d*S* > 1 (Yang et al. 2000). The models were then compared using likelihood ratio tests with 2 degrees of freedom and *P*-values were estimated using the χ^2 approximation (Yang et al. 2000). All *P*-values were then corrected for multiple testing using the FDR method (Benjamini & Hochberg 1995).

Since for each genus tested, a different set of genes may be present, it is possible that any difference in proportion of genes under positive selection between genera could reflect differences in the subset of transcribed genes in each genus, rather than the relative genome-wide proportion of genes under positive selection. To mitigate against this possibility, subsets of each dataset, referred to as paired-ortholog datasets, were made for pairwise comparison of *Senecio* and each comparison genus. In these, only genes assigned to the same *Arabidopsis* proteins in both *Senecio* and the comparison genus were included. Proportions of genes under positive selection in each of these pairs of datasets were then reassessed.

To estimate the strength of purifying selection in each transcript, and subsequently compare this between orthologues of *Senecio* and each comparison genus, the one-ratio (*M0*) model in *PAML* was also applied to the data. This gives an average estimate for d*N*/d*S* across all codons and branches of each transcript. Since most codons are expected to be under purifying selection, the one-ratio estimate of d*N*/d*S* is expected to be dominated by the signal of purifying selection, and lower estimates indicate genes under stronger purifying selection. To quantify the level of parallel functional constraint between *Senecio* and each comparison genus, d*N*/d*S* values were then correlated between pairs of orthologs in each pairwise-ortholog dataset using a Pearson's product moment correlation test.

4.2.5 Functional enrichment analysis

To determine whether genes under selection in *Senecio* were enriched for any specific functional categories, the Gene Ontology (GO) enrichment analysis implemented in *GOrilla* (Eden et al. 2009) was used. First, UniProt accession numbers for *CRB-BLAST* hits for two sets of genes: those under positive selection, the test set, and all other genes with a RBB hit, the reference set, were determined. These test and reference sets were then used as input for the *GOrilla* web server (http://cbl-gorilla.cs.technion.ac.il) in target vs background mode, using *Arabidopsis* as a reference.

4.3 Results

4.3.1 Sequence data processing

Numbers of reads retained after read trimming and subsampling (see methods), ranged from 6.5 to 83 million (Table 4.1). For each genus, the species with the most paired-end reads after trimming was then used for *de novo* transcriptome assembly. These were

Flaveria trinervia, *Glycine max*, *Helianthus bolanderi*, *Linum perenne*, *Oryza sativa*, *Populus tremula*, *Silene vulgaris, Senecio vernalis* and *Solanum melongena.* Numbers of assembled genes ranged from 31,968 (*Flaveria*) to 76,377 (*Silene*) with total lengths ranging from 26,749,542 bp (*Flaveria*) to 82,195,574 bp (*Populus*) and N50 values ranging from 842 bp (*Glycine*) to 1573 bp (*Populus*; Table 4.2). Proportion of reads mapped to the reference transcriptomes for most species were above 70%, although some, particularly in the genus *Linum* as well as *Populus trichocarpa* were significantly lower (Table 4.3).

Table 4.2: Assembly statistics for *de novo* transcriptomes.

Genus	Number of transcripts	Number of transcripts with ORF	Total length (bp)	%GC	N50
Flaveria	31,968	27,741	26,749,542	41.70	1,025
Glycine	59,158	39,762	44,439,668	42.24	842
Helianthus	72,271	51,540	75,926,077	41.09	1,333
Linum	35,628	26,718	42,409,710	44.81	1,555
Oryza	59,047	46,748	67,386,682	49.22	1,526
Populus	46,332	25,730	82,195,574	40.73	1,573
Senecio	39,526	32,848	34,320,951	41.00	1,057
Silene	76,377	52,910	80,794,422	40.61	1,386
Solanum	51,832	39,488	51,483,265	40.40	1,337

Table 4.3: Mapping statistics and heterozygosity.

Genus	Species	Percent mapped	Mean heterozygosity per transcript	Standard deviation heterozygosity per transcript
Flaveria	*anomala*	82.46	0.0126	0.0161
Flaveria	*bidentis*	94.98	0.0035	0.0133
Flaveria	*brownii*	89.99	0.0264	0.0187
Flaveria	*chlorifolia*	83.10	0.0050	0.0132
Flaveria	*pubescens*	92.96	0.0074	0.0160
Flaveria	*ramosissima*	86.23	0.0178	0.0167
Flaveria	*robusta*	78.22	0.0323	0.0213
Flaveria	*trinervia*	96.77	0.0024	0.0124
Glycine	*canescens*	86.60	0.0190	0.0243
Glycine	*clandestina*	81.10	0.0186	0.0227
Glycine	*max*	96.93	0.0125	0.0183
Glycine	*soja*	84.87	0.0227	0.0254
Glycine	*syndetika*	78.48	0.0209	0.0237
Glycine	*tomentella*	72.52	0.0212	0.0231
Helianthus	*annuus*	92.54	0.0088	0.0168
Helianthus	*argophyllus*	95.91	0.0139	0.0242
Helianthus	*bolanderi*	98.13	0.0081	0.0163
Helianthus	*exilis*	92.01	0.0039	0.0088
Helianthus	*petiolaris*	95.67	0.0178	0.0259
Helianthus	*praecox*	91.16	0.0234	0.0269
Helianthus	*debilis*	96.01	0.0211	0.0254
Linum	*bienne*	53.62	0.0223	0.0208
Linum	*grandiflorum*	60.80	0.0122	0.0184
Linum	*hirsutum*	46.24	0.0102	0.0217
Linum	*lewisii*	93.60	0.0043	0.0139
Linum	*macraei*	44.59	0.0162	0.0159
Linum	*perenne*	98.05	0.0058	0.0113
Linum	*tenuifolium*	37.60	0.0034	0.0132
Linum	*usitatissimum*	69.98	0.0245	0.0205
Oryza	*barthii*	95.56	0.0031	0.0130
Oryza	*glumipatula*	94.76	0.0034	0.0135
Oryza	*granulata*	75.57	0.0053	0.0145
Oryza	*meyeriana*	81.32	0.0304	0.0248
Oryza	*nivara*	91.39	0.0033	0.0148
Oryza	*officinalis*	45.13	0.0042	0.0165

Table 4.3 cont.

Oryza	*rufipogon*	81.31	0.0305	0.0249
Oryza	*sativa*	98.97	0.0030	0.0129
Populus	*euphratica*	93.49	0.0092	0.0168
Populus	*pruinosa*	86.37	0.0114	0.0217
Populus	*tomentosa*	93.49	0.0183	0.0204
Populus	*tremula*	97.91	0.0080	0.0154
Populus	*tremuloides*	94.03	0.0069	0.0142
Populus	*trichocarpa*	11.26	0.0075	0.0178
Senecio	*aethnensis*	90.87	0.0091	0.0179
Senecio	*chrysanthemifolius*	91.37	0.0075	0.0159
Senecio	*madagascariensis*	92.01	0.0136	0.0204
Senecio	*flavus*	89.33	0.0045	0.0164
Senecio	*gallicus*	93.10	0.0091	0.0147
Senecio	*glaucus*	96.00	0.0115	0.0165
Senecio	*leucanthemifolius*	93.68	0.0102	0.0183
Senecio	*vernalis*	95.62	0.0073	0.0143
Silene	*acaulis*	85.05	0.0127	0.0246
Silene	*diclinis*	92.73	0.0110	0.0240
Silene	*dioica*	88.28	0.0146	0.0245
Silene	*gallica*	85.90	0.0057	0.0193
Silene	*vulgaris*	98.21	0.0147	0.0206
Silene	*latifolia*	88.55	0.0121	0.0220
Solanum	*berthaultii*	66.61	0.0115	0.0140
Solanum	*habrochaites*	74.75	0.0034	0.0128
Solanum	*incanum*	78.70	0.0051	0.0154
Solanum	*melongena*	97.73	0.0056	0.0189
Solanum	*pennellii*	81.84	0.0037	0.0136
Solanum	*phureja*	73.76	0.0068	0.0192
Solanum	*torvum*	88.60	0.0048	0.0157
Solanum	*dulcamara*	75.29	0.0062	0.0153

4.3.2 Phylogenetic analysis

The 4-fold site and non-coding region datasets used for phylogenetic reconstruction included between 2.4 million and 5.2 million sites, with proportions of missing data between 37% and 74% (Table 4.4). For all genera, highly supported species phylogenies

were reconstructed with only two nodes amongst the whole set of species trees having bootstrap support below 95% (Fig. 4.1). Total lengths of trees varied between 0.34 and 1.01 substitutions per site (Table 4.4). The phylogeny for *Senecio* was identical to the species topology estimated in Chapter 3 (using the same dataset, but a different analytical pipeline).

Table 4.4: Phylogenetic reconstruction statistics.

Genus	Number of sites in neutral datasets (bp)	Proportion of missing data in neutral datasets	Total tree length (substitutions per site)
Flaveria	3,921,687	0.37	0.132
Glycine	2,764,446	0.49	0.034
Helianthus	5,068,526	0.50	0.056
Linum	3,637,606	0.74	1.011
Oryza	5,957,007	0.55	0.330
Populus	2,480,919	0.57	0.883
Senecio	3,992,573	0.47	0.202
Silene	3,853,911	0.60	0.212
Solanum	5,245,254	0.59	0.255

Table 4.5: Phylogenetic incongruence in each genus.

Genus	% genes rejecting species tree in SH test before FDR ($P < 0.05$)	% genes rejecting species tree in SH test after FDR ($Q < 0.05$)
Flaveria	5.80	0.32
Glycine	2.60	0.51
Helianthus	7.31	0.31
Linum	0.67	0.19
Oryza	19.74	0.53
Populus	2.82	0.60
Senecio	5.92	0.29
Silene	3.51	0.08
Solanum	0.82	0.27

Percent of genes significantly rejecting the species tree ($Q < 0.05$) for each genus ranged from 0.67 in *Linum* to 19.24 in *Oryza* however these figures were all below 1% after FDR correction (Table 4.5). Since these SH test proportions are estimated after other filters have been applied (d*S* and heterozygosity outlier filters) they may be underestimates of the genome-wide proportion of phylogenetic incongruence.

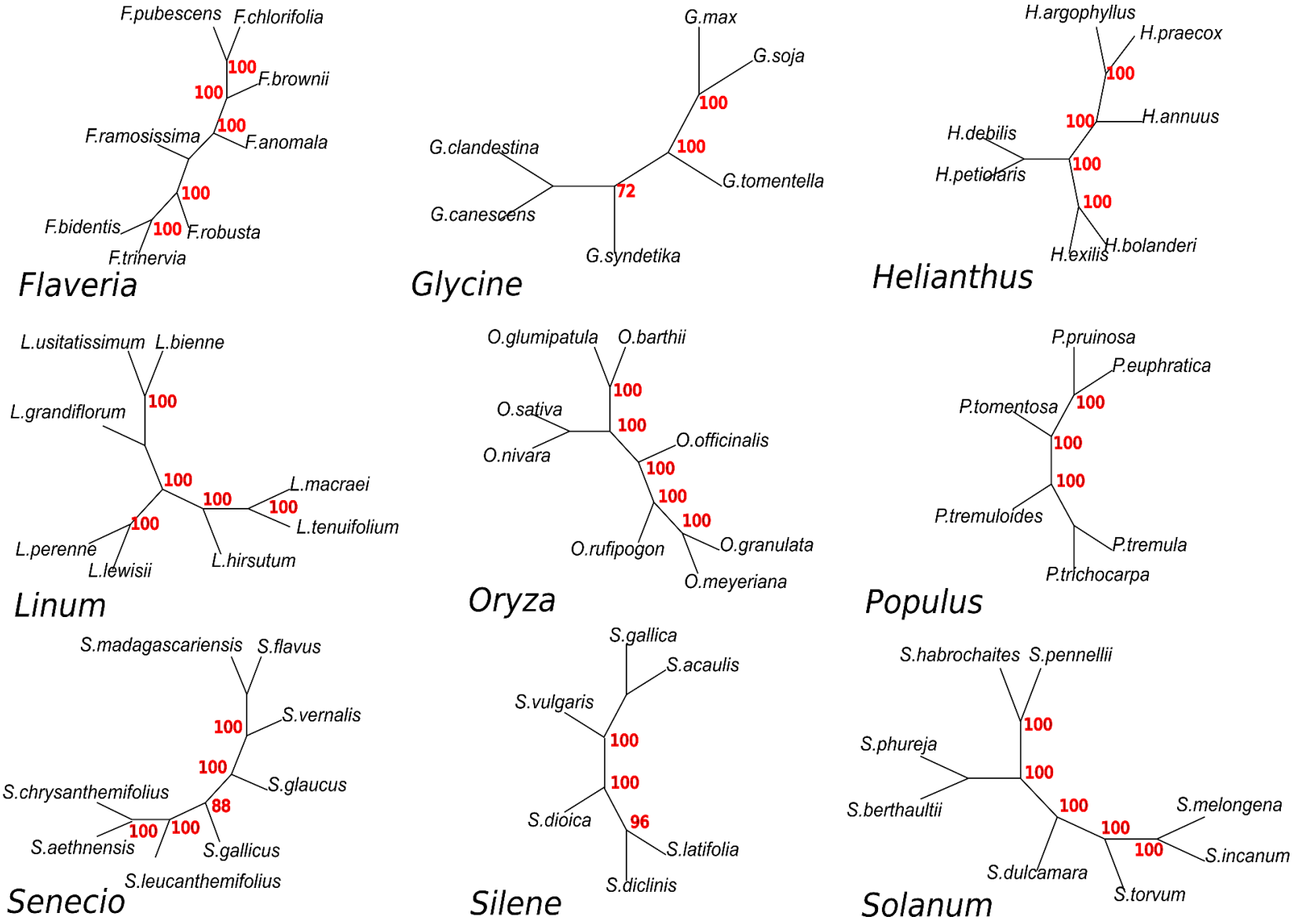

Figure 4.1: Estimated species topology for each genus tested. Branch lengths are arbitrary and don't reflect distance and node labels in red are bootstrap support.

4.3.3 Dataset filtering

After applying filters to each CDS-dataset, between 678 (*Populus*) and 7696 (*Helianthus*) genes remained for the tests of positive selection (Fig. 4.2). All genera except *Populus* retained over 2000 genes. The missing data filter (<5 species) was responsible for the removal of by far the most genes. Heterozygosity is likely to vary between species within each genus, so genes which had high heterozygosity in any species were removed. While this removes a reasonably high proportion of genes, it is expected to significantly reduce the chance of incorrect mapping of paralogues. Overall, this approach applies a high level of stringency, with multiple filters addressing multiple potential sources of bias, while still retaining a large enough number of genes to estimate the frequency of positive selection on a genome-wide scale.

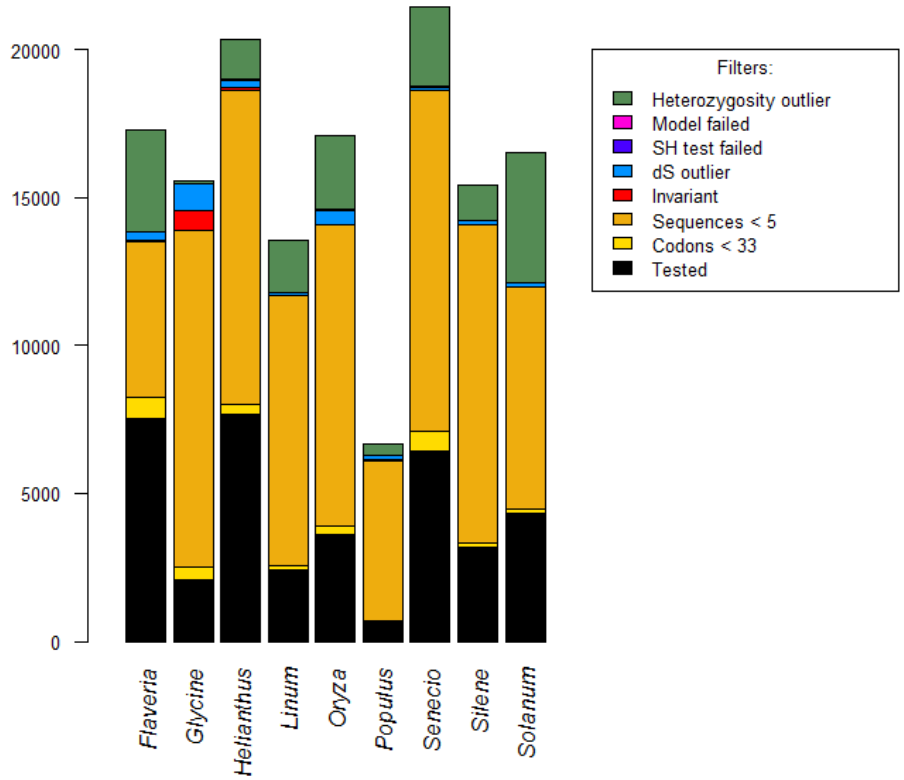

Figure 4.2: Numbers of genes used in the analyses of positive selection for both *Senecio* and all comparison genera. Numbers used in the analysis are shown in black and those failing each filter are shown in various colours.

4.3.4 Genes under positive selection in *Senecio*

Following filtering of genes and correction for multiple testing, 23 genes showed evidence of positive selection in *Senecio* based on the likelihood ratio tests comparing models *M7* and *M8* in PAML. These were associated with a wide variety of functions (Table 4.6). The functional enrichment test failed to find any GO terms significantly enriched in the positively selected subset of genes after FDR correction, although one term, signalling receptor activity (GO:0038023) just fell short of significance following (FDR corrected *P*-value: 0.067).

Table 4.6: Genes under positive selection in *Senecio.*

N codons	Likelihood ratio	*Q*-value	UniProt accession	Gene descriptionGene description
616	17.64	0.047	O82333	Probable indole-3-acetic acid-amido synthetase GH3.1
211	20.19	0.022	Q9FPJ5	Leucine-rich repeat (LRR) family protein
216	17.49	0.047	-	
615	18.48	0.035	Q94F62	BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1
534	22.94	0.011	A9LLI7	Exosome complex exonuclease RRP6
439	28.02	0.002	Q8VZC0	UDP-glucuronic acid decarboxylase 1
522	32.27	0.001	-	
313	20.17	0.022	-	
694	21.75	0.017	Q9SIF2	Heat Shock Protein 90
160	19.26	0.028	-	
553	19.00	0.028	Q9LV91	4-alpha-glucanotransferase DPE1
1139	17.97	0.042	B3H739	ketose-bisphosphate aldolase class-II family protein
866	21.45	0.017	-	
151	17.40	0.047	Q9ZQX4	V-type proton ATPase subunit F
224	27.51	0.002	Q9FFB3	Vacuolar protein sorting-associated protein 24 homolog 1
275	19.16	0.028	-	
199	23.17	0.011	-	
125	21.02	0.018	Q9LTP9	Phosphorylated adapter RNA export protein
404	19.81	0.025	Q93VP4	Transcription factor IIA, alpha/beta subunit
422	17.37	0.047	Q9FN03	Ultraviolet-B receptor UVR8
118	21.26	0.017	F4JG25	Uncharacterized protein
309	19.12	0.028	Q8RXT5	Bromo-adjacent homology (BAH) domain-containing protein
628	22.92	0.011	Q9C950	Paf1C-like transcription factor

4.3.5 Rates of selection in *Senecio* relative to other plant genera

The dearth of comparative studies identifying d*N*/d*S* based signatures of positive selection on a genome-wide scale means it is unclear whether the number of genes under positive selection in *Senecio* is high or low relative to plants in general. To put the rates of positive selection in *Senecio* in the context of plants more generally, an identical analysis was conducted on eight comparison genera. The results showed that *Senecio* had the second highest rate of positive selection of the nine genera tested. Only in *Helianthus* were more genes, and a higher proportion of all those tested, under positive selection (Fig. 4.3). Only considering genes within the pairwise-ortholog datasets, showed that these comparisons are robust to differences in the subset of genes analysed. For comparisons for which *Senecio* had a higher rate of positive selection in the total CDS dataset, *Senecio* also had a higher proportion of genes under selection in respective the pairwise-ortholog dataset, and the single comparison for which *Senecio* was lower, also showed the same pattern in the pairwise- ortholog dataset (Fig. 4.4). The correlation between these proportion of genes under positive selection in all comparative genera in the paired-ortholog and total CDS datasets was highly significant and strongly positive (Pearson's product moment correlation; $R = 0.97$; $P = 5.02 \times 10^{-5}$). The power of the *PAML* likelihood ratio tests can be increased by two important variables, number of species and species divergence time (Anisimova et al. 2001). However, neither two proxies for divergence time (mean d*S* per genus, total tree length from the neutral dataset) nor the number of species per genus were significantly correlated with proportion of genes under positive selection (Pearson's product-moment correlation tests: d*S* vs selection: $P = 0.36$, $R = -0.345$; tree length vs selection: $P = 0.34$, $R = -0.359$; Number of species vs selection: $P = 0.98$; $R = -0.009$)

suggesting that the differences in rates of selection between genera are not dependent on differences in power resulting from variable divergence time and species number.

Finally, the association between the selective regime at orthologues between *Senecio* and other taxa was investigated. No genes under positive selection were shared between *Senecio* and any of the comparison genera. However, when d*N*/d*S* (from the one-ratio *M0* model) were compared between *Senecio* and each comparison genus, all comparisons were significantly correlated, and correlation coefficients ranged from 0.15 to 0.43, suggesting that levels of functional constraint among many genes are surprisingly constant, even across vast phylogenetic distances (Fig. 4.5).

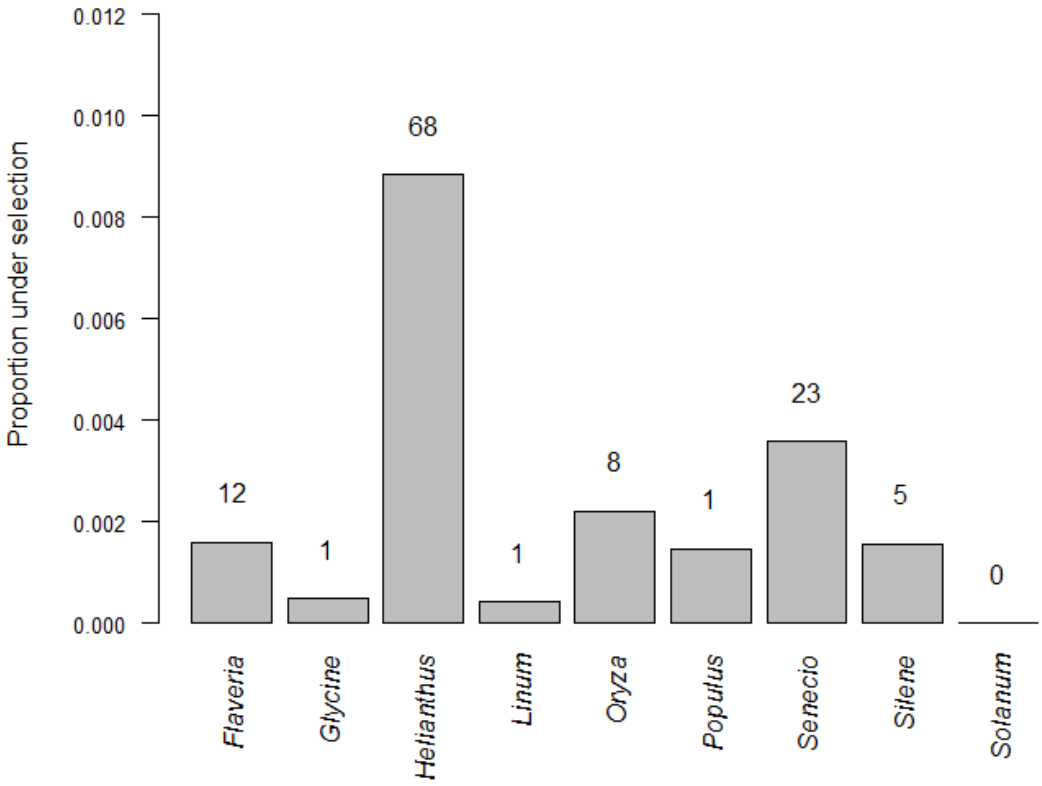

Figure 4.3. Proportions of genes under positive selection in the total tested datasets of *Senecio* and all comparison genera. Bar heights represent the proportion of genes tested for each dataset and numbers above bars are the number of genes under selection.

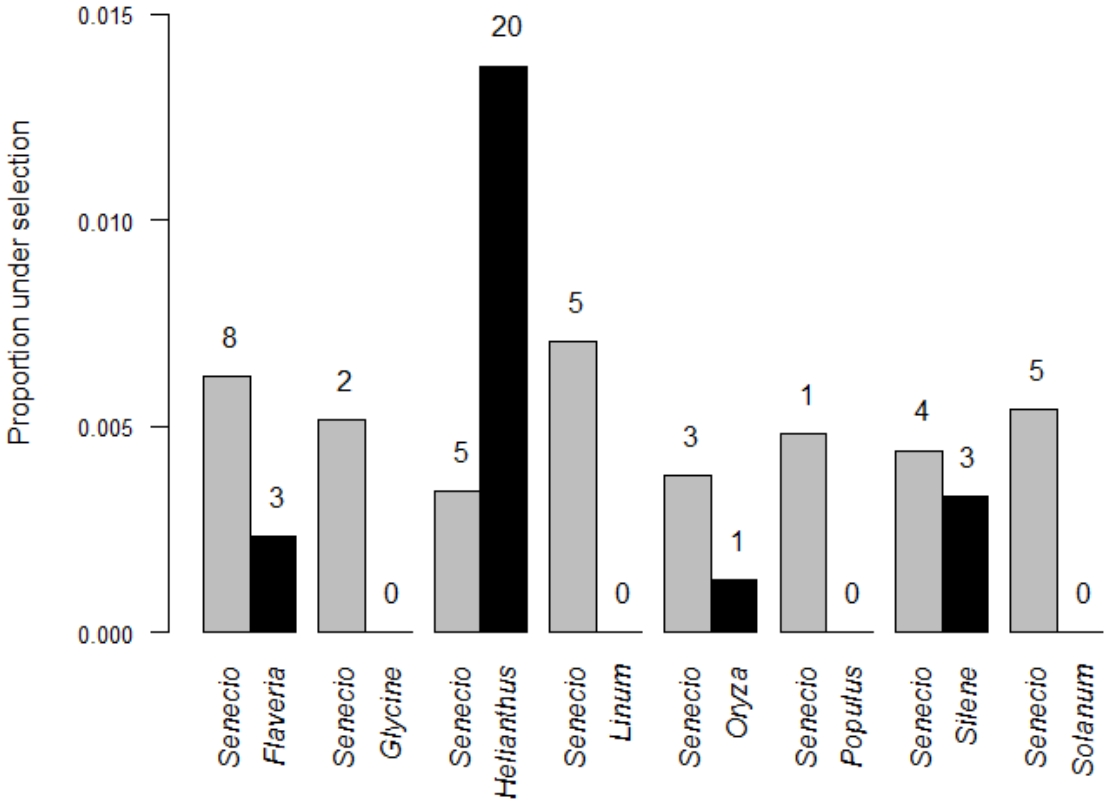

Figure 4.4. Proportions of genes under positive selection in the pairwise orthologue datasets of *Senecio* and each comparison genus. The pairwise orthologue dataset for each comparison genus is the subset of genes in *Senecio* and the comparison genus which share the same Conditional Reciprocal Best *BLAST* hits to the *Arabidopsis* proteome. Bar heights represent the proportion of genes tested for each dataset and numbers above bars are the number of genes under selection. Pairs of bars show results for each pairwise-ortholog dataset with *Senecio* in grey and comparison genera in black.

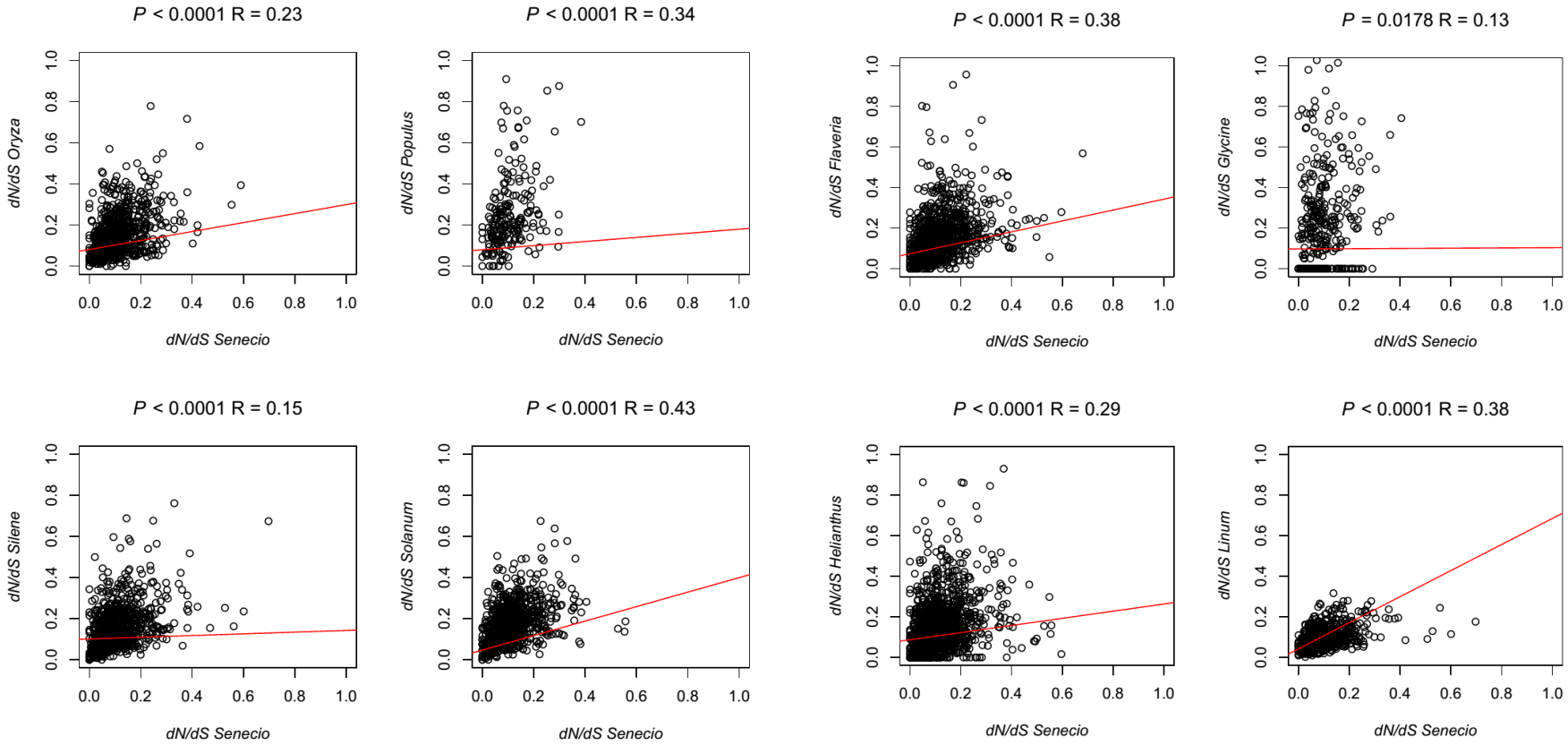

Figure 4.5: Correlations of d*N*/d*S* in orthologous genes between *Senecio* and each comparison genus. Red lines show linear regressions for each test.

4.4 Discussion

4.4.1 Validity of results

Codon-based tests of positive selection offer a powerful approach to identifying the genomic targets of positive selection and comparing the frequency of positive selection between taxa, although it has only rarely been used in such a comparative framework (but see for example Kane et al. 2011; Kapralov et al. 2013). However, there are several potential pitfalls and sources of bias to be avoided using these methods. This is especially true when comparing results from multiple large datasets from different sources, as has been done here. Therefore in this chapter, a conservative approach was taken at all stages of the analysis.

Sequencing and bioinformatic error are considerable potential sources of false positives in likelihood based tests of selection (Fletcher & Yang 2010; Jordan & Goldman 2012; Mallick et al. 2009). The mapping and SNP calling pipeline implemented here applies every available option to reduce the inference of incorrect SNPs. Both mapping and base quality thresholds are high, and realignment of mapped reads around indels reduces the potential for erroneous SNPs due to indel misalignment (DePristo et al. 2011). Even with these safeguards in place, it may be possible for erroneous SNPs to slip through the net, especially as a consequence of misalignment of paralogous sequences (Malhis & Jones 2010). Therefore three further filters were used to remove genes with a high probability of paralogous misalignment, two based on SNP density (removing synonymous divergence and heterozygosity outliers) and one phylogenetic (SH tests; Shimodaira & Hasegawa 2001). A further source of bias - incorrect tree topology due for example to hybridisation, was also ameliorated by the application of SH tests. However, both Yang et al. (2000) and Pie (2006) found the effect of incorrect topology to have only a minor effect on the detection of positive selection using these methods, so this

source of bias is expected to be relatively unimportant compared to bioinformatics concerns. Finally, all *P*-values were corrected for multiple testing using the method of Benjamini & Hochberg (1995). The multiple filters and precautions implemented at every stage of the analysis undoubtedly remove a high proportion of potentially testable genes (Fig. 4.2), some of which may genuinely be under positive selection. Nevertheless, a conservative approach such as this is essential for the robust identification of the targets of selection, and especially in comparison between taxa.

4.4.2 A diverse collection of protein-coding genes are under selection in *Senecio*

The results show strong evidence for positive selection on the amino acid sequences of at least 23 genes in *Senecio.* Many researchers have argued for the unimportance of amino-acid changes in adaptive evolution, and suggested that adaptive substitutions are overwhelmingly likely to occur in *cis*-regulatory regions affecting expression (Carroll 2000; Prud'homme et al. 2007; Wray 2007). The main rationale behind this view is that cis-regulatory mutations are likely to be less pleiotropic (Wray 2007). This thesis does not examine expression divergence, or provide any comparison between the two types of variation, but the fact that evidence for positive selection was found in all genera except one, certainly suggests that protein-coding divergence is not an insignificant component of adaptive evolution in plants.

In terms of the identity of the genes under selection in *Senecio*, no obvious trend was found. While one GO term was close to significance, Signalling Receptor Activity is such a broad term, covering such a wide array of biological processes that deriving a biological meaning from this would be challenging even if it was significant. Nevertheless, the identity of some genes under selection deserve mention. For example *Paf1C-like transcription factor* (Q9C950) is required for the vernalisation response in *Arabidopsis*, in which prolonged cold during winter is

required for flowering to occur (Oh et al. 2004). Since the winter temperatures in the native ranges of the species included in the analysis vary hugely, it is conceivable that temperature adaptation could be responsible for selection on this gene. *Ultraviolet-B receptor UVR8* (Q9FN03) is known to control UV protection responses in plants (Brown et al. 2005). Levels of UV radiation vary with both altitude and cloud cover, both of which vary greatly between the ranges of the species examined, so it is possible that differing UV protection needs amongst the species could have facilitated strong selection at this gene. Nevertheless, making any, even tentative, conclusions based on the identity of the *Arabidopsis* homologs of the genes under selection in *Senecio* would be premature. These genes should, however, be viewed as interesting candidates for future studies examining the genomic basis for adaptation in the genus.

4.4.3 Selective regimes in orthologues are highly correlated in different plant taxa

The comparison of d*N*/d*S* in orthologues between *Senecio* and each comparison genus found strikingly strong correlations in every comparison. Since the overwhelming majority of genes have a value of d*N*/d*S* well below one, and molecular evolution in coding sequences is dominated by purifying selection in general (Zhai et al. 2012) this is likely a result of similar levels of functional constraint (purifying selection) acting in orthologues across plants. Kane et al. (2011) found similar correlations in a study comparing the evolution of orthologues within plant families. The results in this chapter show that the parallel levels of gene-specific functional constraint extend to much higher taxonomic levels, given that the last common ancestor of *Oryza* and *Senecio* was thought to have existed around 140 to 150 million years ago (Chaw et al. 2004). This underlines the fact that the evolution of protein-coding genes is overwhelmingly dominated by purifying selection, and that many of the same genes are likely

to remain under the strongest levels of functional constraint, even over hundreds of millions of years of plant evolution.

In sharp contrast to this, there was no overlap between any genes under positive selection in any of the genera. This underscores that molecular evolution may be predictable in terms of the genes under purifying selection, but less so with regard to the genes involved in adaptive evolution. Just as we can predict that non-synonymous sites are likely to evolve more slowly than synonymous sites on the whole, a substantial portion of genes within the genome are likely to be under stronger purifying selection and evolve particularly slowly at the amino acid level. However, where in the genome the positive selection which is responsible for adaptive evolution is likely to occur, appears to be far less congruent between taxa. Nevertheless, some examples of convergent adaptive evolution at the molecular level are known. For example, many genes have undergone convergent evolutionary changes at the molecular level during the evolution of echolocation in cetaceans and bats (Parker et al. 2013). The lack of any such convergent evolution in the plant genera analysed in this chapter may simply reflect the very different selection pressures the taxa have faced during their evolution. Alternatively, convergent positive selection between disparate taxa may truly be incredibly rare. As genomic-scale data becomes available for more and more species, general patterns in the predictability of genetic evolution such as these are likely to become clear (Stern 2013).

4.4.4 *Senecio* has a high rate of positive selection relative to many plant genera

The results in this chapter show the first evidence that *Senecio* has a high rate of positive selection relative to plants in general. The other genus which stood out was *Helianthus,* and while *Senecio* had a higher proportion of genes under positive selection than seven of the

eight comparison genera, *Helianthus* showed evidence for positive selection in more than twice the fraction of genes as *Senecio.* Interestingly, Gossmann et al. (2010) also found high rates of positive selection in *Helianthus* relative to other plants. The meaning and significance of this finding, however, is not easy to interpret. Firstly, it is important to point out, that the proportion of genes under selection is a measure of the breadth of selection across the genome, rather than its strength per se. The number of distinct phenotypes under selection in each species, the number of genes which can facilitate this adaptive phenotypic divergence and the number of favourable mutations available would also be expected to affect this number in addition to the strength of selection. Secondly, the tests implemented in Chapter 4 only detect a specific type of positive selection, selection on amino acid substitutions within proteins (Yang et al. 2000). Other sources of phenotypic variation are also likely to be under selection and to have contributed to adaptation in the species, particularly cis-regulatory mutations which cause expression variation.

Nevertheless, it seems likely given the results that some fundamental differences exist between taxa in terms of positive selection at the genomic level. This begs the question of what the factors are that influence the frequency of positive selection during clade diversification? One factor which theory predicts should have a substantial effect on the rate of positive selection is effective population size (Kimura 1983). This is because the rate of adaptive evolution is expected to be proportional to $\mu N_e s$ where μ is the mutation rate, N_e is the effective population size, and s is the strength of selection. This has been borne out in empirical studies, which have found correlations between measures of selection and effective population size (Gossmann et al. 2012, 2010; Strasburg et al. 2011). However, since the methods used in this chapter detect selection in a phylogenetic context, across multiple

species which may have differing population sizes, any genus-specific estimate of population size would be unlikely to be meaningful. Furthermore, given that effective population size may have changed drastically over the course of the divergence of these taxa, and the tests of selection cannot date when the selection occurred, any correlation in the context of this study would be additionally complicated.

It might be expected that taxa which have adapted to more diverse ecological niches would be under stronger positive selection, and would thus have a greater rate of adaptive evolution at the molecular level. Indeed, Kapralov et al. (2013) found that *Schiedea*, an adaptive radiation of Hawaiian plants which has diverged into a wealth of different habitats and growth forms, had higher rates of positive selection than many mainland taxa. The study found positive selection in 13.89% of genes tested, although only 36 genes were tested overall. Other similar studies are sparse, and more importantly, the level of ecological divergence within a taxon is almost impossible to quantify in an unbiased way. Ecological niches are by their nature multidimensional, and environmental factors which may be important in one clade may be irrelevant in another, making comparison of levels of niche differentiation between taxa highly challenging.

It is possible that either relatively large effective population sizes or strong selection during the diversification of *Senecio* could have contributed to the relatively high rates of positive selection observed here. The species analysed certainly appear to be highly ecologically diverse. Future work which could help to resolve this, could include detailed measurements of many ecological factors across the ranges of these species, allowing ecological niche models to be estimated and niche divergence to be quantified (Warren et al. 2010).

Furthermore, ecological experiments such as reciprocal transplant experiments, could quantify the level of adaptive divergence in these species in wild conditions.

4.4.5 Do plants experience high levels of positive selection in general?

It has been suggested that plants might fundamentally differ from animals in the strength and breadth of selection across their genomes. For instance, many more genes are expressed in the haploid phase of the plant life-cycle (the gametophyte) than the equivalent phase in animals (eggs and sperm; Borg et al. 2009). This may lead to more efficient selection in these genes in plants because they cannot be masked by another allele when expressed in the haploid state (Chibalina & Filatov 2011). On the other hand, Gossmann et al. (2010) found that plants had lower rates of adaptive evolution than animals, and suggested that this difference may be related to differences in average population sizes between the two kingdoms. The analyses in this chapter represent the largest study to date to determine the breadth of positive selection in plants and found evidence of positive selection in all but one genus (*Solanum*). Since Gossmann et al. (2010) use a different approach which detects selection on a different timescale and has different sensitivities, the their results and those in this chapter are not comparable. Similar (although certainly not identical) analytical pipelines to those here have been applied in *Drosophila* and mammals (Larracuente et al. 2008; Kosiol et al. 2008). These studies found 10.32% and 2.42% of genes to have been under selection respectively, far higher than any of the plant genera analysed here. This would appear to support the results of Gossman (2010) that plants have relatively low rates of positive selection. However, since the clades in these two studies are far older than those analysed in this study, far more time has elapsed in which positive selection could have happened (Larracuente et al. 2008; Kosiol et al. 2008). Therefore, no definitive conclusions can be drawn

at this stage. Further analyses, which compared similar (in terms of divergence time) taxa in plants and animals would significantly improve current understanding of potential differences in the role of selection between the major groups of life.

4.4.6 Conclusions

The results in this chapter identify several genes which are likely to have been under positive selection in *Senecio*, which should be key targets of future studies. Furthermore, they have shown both that rates of positive selection are high in *Senecio,* and are highly variable among plant genera more generally. In contrast to these findings, the correlation of the strength of purifying selection between *Senecio* genes and their orthologues in other species, underline the fact that while a small proportion of the genome may be involved in adaptive evolution, the majority of genes are under strong functional constraint which remains constant over vast swathes of evolutionary time.

Chapter 5: General discussion

5.1 *Senecio aethnensis* and *Senecio chrysanthemifolius* - an example of ecological speciation with gene flow?

Darwin's original concept of speciation highlighted the importance of diversifying selection as the major force responsible for the origin of species (Darwin 1859; Pinho & Hey 2010). While this was emphasised less during much of the 20th century, recent work on ecological speciation has brought selection back to the forefront of speciation research (Rundle & Nosil

2005; Nosil et al. 2009, 2005). Further clear examples of recent and ongoing speciation are needed, however.

The habitats of *Senecio aethnensis* and *S. chrysanthemifolius*, despite occurring only a few kilometres apart are vastly different, varying in temperature, UV exposure and many other ecologically important variables (James & Abbott 2005). This extreme difference in habitat is mirrored by substantial phenotypic divergence in trait such as leaf shape, inflorescence size, flowering time, seed germination temperature, stature and a multitude of other traits (Hegarty et al. 2009; Brennan et al. 2012; Ross 2010). However they hybridise readily, and form a stable hybrid zone (Brennan et al. 2009). In terms of geography, *S. aethnensis* is endemic to the upper slopes of Mount Etna, whereas *S. chrysanthemifolius* is found only on the lower slopes of Mount Etna and its surroundings. There is also both phenotypic (Brennan et al., 2009; Ross et al., 2010), and molecular (Brennan et al., 2009; Chapman et al., 2013; Muir et al., 2013) evidence for divergent selection between the species. Furthermore, weak post-zygotic reproductive isolation exists between the species (Brennan et al. 2014; Chapman et al. 2016). These results suggest that they may be an example of ecological speciation with gene flow, possibly diverging as a result of ecological selection imposed by the growth of Mount Etna. The results in this thesis make several important contributions towards bolstering this view.

Firstly, the results provide estimates of the demographic parameters of the species divergence. Results in Chapter 2, confirm that a model of divergence with gene flow is a significantly better fit to the data than one without. Furthermore, the estimates of divergence time between the species fits strikingly well with the growth of Mount Etna. While no causal relationship can be absolutely established from this, it is certainly strong circumstantial

evidence that the growth of Mount Etna may have been pivotal in the species' divergence. Secondly, the analyses in Chapter 3 find that *S. aethnensis* and *S. chrysanthemifolius* are sister species. As discussed in Chapter 3, the species sampling is not exhaustive, so it is still possible that this is not the case. However, a finding that they were not sister species would certainly have debunked the hypothesis of ecological speciation with gene flow, and made a scenario of secondary contact more likely, so this result is important. Obtaining similar datasets for all species within the clade should be a major goal of future research.

Some further results supporting ecological speciation with gene flow between the species are detailed in my further work which was not included in this thesis (Muir et al. 2013; Filatov et al. 2016). Muir et al. (2013) details demographic analyses which broadly corroborate the results of Chapter 2 and also estimate the rate of gene flow, finding it to be above one effective migrant per generation in both directions. This high rate of gene flow would be high enough, assuming migration-drift equilibrium and in the absence of divergent selection, to prevent the initiation and maintenance of species divergence (Wright 1931). The fairly modest difference in divergence time estimate from those in Chapter 2 and Muir et al (2013) are likely to be an artefact of the type of data used, since they included non-synonymous sites at which any estimate of mutation rate (used to convert parameter estimates into demographic units) are likely to be unreliable. Thus the divergence time estimates in Chapter 2, based on four-fold degenerate sites are likely to be far more reliable. It also provides evidence that differentially expressed genes are under stronger divergent selection than those with similar expression levels between the species. These two results together support the argument that the species have diverged under divergent selection and suggest that expression divergence may have played a role in their adaptation (Muir et al. 2013).

Filatov and colleagues (2016) provide two further crucial pieces of evidence to support the ecological speciation with gene flow scenario. Firstly, it addresses concerns that previous estimates of demographic parameters (Osborne et al. 2013; Muir et al. 2013; Chapman et al. 2013) could have been influenced by the individuals sampled not being from the absolute extremes of the altitudinal gradient on Mount Etna (Abbott & Brennan 2014). By analysing data from populations sampled from the altitudinal extremes of the species ranges, as well as from similar altitudes as in previous studies (Osborne et al. 2013; Muir et al. 2013; Chapman et al. 2013), the results suggest that while gene flow may have been overestimated by the choice of populations in Chapman et al. (2013), divergence time estimates were robust to differences in population sampling. Secondly, the study provides the first test comparing a model of continuous gene flow since the species split, with one of only recent gene flow. This is important since previous tests (including those in this thesis) are unable to differentiate between scenarios of continuous gene flow since speciation and those of speciation without gene flow followed by secondary contact. There was no evidence in favour of the model of only recent gene flow, suggesting that a secondary contact scenario is unlikely.

Assuming, as now seems probable, that the species are an example of recent ecological speciation with gene flow, the system has many unique qualities making it an indispensable model for speciation research. For instance, much of the recent discourse around ecological speciation has concerned the existence and nature of highly differentiated genomic "islands of speciation" responsible for maintaining species differences in the face of gene flow. Interpretation of genomic patterns of differentiation purported to support this model have been contentious, and some authors have argued that highly differentiated regions could result from selection following divergence, rather than a local reduction in gene flow caused

by divergent selection (Noor & Bennett 2010; Turner & Hahn 2010; Cruickshank & Hahn 2014). The strength of *S. aethnensis* and *S. chrysanthemifolius* to approach this question, is their known contemporary hybridisation, and that this hybridisation occurs across a steep ecological cline which is likely to confer strong divergent selection. By sampling across the cline, geographic and genomic cline based approaches can be used which measure gene flow at individual loci far more directly than by statistics such as *Fst*, which are also affected by factors other than gene flow (Barton & Hewitt 1985; Gompert & Alex 2011).

The steep altitudinal cline also renders the ecological factors affecting the species' divergence far easier to fathom than in many species. While it is certainly possible that many factors, such as soil composition and biotic interactions, may vary across the species' ranges in a way which does not correlate with elevation, altitude is almost unique as an ecological measure, in that it universally correlates with several factors; temperature, solar radiation, proportion of UV-B radiation and partial pressure of atmospheric gases; likely to have profound biological significance (Korner, 2007). Thus a simple measure of the altitude at which a plant was collected is likely to serve as a proxy for a large proportion of the ecological factors affecting the species divergence.

Linking this ecological divergence to the emergence of reproductive isolation in the species will be an important next step. Recent work has found evidence of a low level of intrinsic reproductive isolation between them (Brennan et al. 2014; Chapman et al. 2016) and determining whether the loci involved in this incipient reproductive isolation have evolved as a direct consequence of ecological divergence is crucial. For example, if regions of the genome containing genomic incompatibilities between the species also contain regions controlling important ecological traits, or show genomic signatures of divergent selection, then

ecological divergence would be directly implicated, and a genetic mechanism linking ecological divergence with reproductive isolation would be shown. A very recent study shows evidence for this, with QTLs for hybrid breakdown overlapping those for key morphological traits in the species, suggesting that ecological divergence and the evolution of RI may be linked (Chapman et al. 2016). The significance of the results in this thesis, in substantially bolstering the case for ecological speciation in the species, lay the foundations for future research in the species, and the further development of *S. aethnensis* and *S. chrysanthemifolius* into a major study system for speciation biology.

5.2 The Mediterranean *Senecio* species-complex as a model for diversification with gene flow

The botanist Edgar Anderson was an early proponent of the evolutionary importance of hybridisation when it was generally viewed as unimportant (Anderson 1949). The subsequent sea-change in opinion in his favour is illustrated in the following quote from Warren H. Wagner "We used to make fun of Edgar Anderson by saying that he was finding hybrids under every bush. Then we realized that even the bushes were hybrids" (Abbott et al. 2013). Hybridisation is increasingly recognised as a crucial factor in adaptation and speciation (Seehausen 2004; Whitney et al. 2010; Morjan & Rieseberg 2004). It can have unpredictable consequences, either aiding or countering adaptation, and either generating new species or merging existing ones (Seehausen 2004; Webb et al. 2011; Abbott et al. 2013; Kleindorfer et al. 2014). What remains particularly unclear is to what extent the diverse potential outcomes of hybridisation are predictable (Abbott et al. 2013). For instance, what is the distribution of fitness effects of introgressed genes? How does the relatedness of hybridising taxa affect the

potential for negative or adaptive consequences of hybridisation? And what proportion of hybridisation events are likely to lead to speciation?

The Mediterranean *Senecio* species-complex has been suggested as a useful system to study such questions (Comes & Abbott 2001). The results of Chapter 3 identify multiple previously unknown examples of hybridisation and introgression between the species and underscore the ubiquity of interspecific gene exchange in *Senecio*. Combining the previous known examples of hybrid speciation and introgression amongst the species in the Mediterranean *Senecio* species complex with the novel ones shown in this thesis, suggest that at least 13 of approximately 26 species in the clade (Comes & Abbott 2001) are likely to have been involved in interspecific hybridisation to some extent (Table 5.1). Judging by the increased ability to detect introgression using high-throughput datasets in Chapter 3, relative to previous studies using smaller datasets (Comes & Abbott 2001), it seems likely that this may increase as genomic-scale data becomes available for more species in the clade. Furthermore, they contain a huge variety of types of hybridisation outcomes, including as they do, examples of both homoploid and allopolyploid hybrid speciation, adaptive and (presumably) non-adaptive introgression (Table 5.1). Thus the species represent a fascinating system in which to study the various consequences of hybridisation in a natural system. The further development of genomic resources in the clade, especially the completion of the *S. squalidus* genome (T. Batstone, M.A. Chapman, O.G. Osborne, D.A. Filatov, R.J. Abbott, and S.J. Hiscock, in preparation) and generation of data for species not covered in this thesis will allow the detection of the specific loci which have introgressed between the species, allow confirmation of hypotheses of hybridisation between the species, and may well suggest previously unknown instances of hybridisation, as in Chapter 3. Thus, the future of the Mediterranean

Senecio species complex as a model for hybridisation and its evolutionary consequences is bright.

Table 5.1: Summary of known episodes of hybrid speciation and introgression within the Mediterranean *Senecio* species complex.

	Description	Reference
Hybrid speciation	Allopolyploid origin of hexaploid *S. cambrensis* from hybridisation between diploid *S. squalidus* and tetraploid *S. vulgaris*	(Lowe & Abbott 1996)
	Allopolyploid origin of tetraploid *S. eboracensis* from hybridisation between between diploid *S. squalidus* and tetraploid *S. vulgaris*	(Lowe & Abbott 2004)
	Homoploid hybrid origin of diploid *S. squalidus* from hybridisation between *S. aethnensis* and *S. chrysanthemifolius* and subsequent transplantation to the UK	(James & Abbott 2005)
	Possible allopolyploid origin of tetraploid *S. vulgaris* from hybridisation between *S. vernalis* and another unknown species	(Ashton & Abbott 1992)
	Allopolyploid origin of hexaploid *S. hoggariensis* from hybridisation between between diploid *S. flavus* and *S. glaucus*	(Kadereit et al. 2006)
	Allopolyploid origin of hexaploid *S. teneriffae* from hybridisation between diploid *S. glaucus* and tetraploid *S. vulgaris*	(Lowe & Abbott 1996)
	Allopolyploid origin of tetraploid *S. mohavensis ssp. breviflorus* from hybridisation between diploid *S. flavus* and *S. glaucus*	(Kadereit et al. 2006)
Introgressive hybridisation	Origin of *S. vulgaris var. hibernicus* through introgression from *S. squalidus* to *S. vulgaris*	(Kim et al. 2008)
	Introgression across the stable hybrid zone between *S. aethnensis* and *S. chrysanthemifolius*	(James & Abbott 2005) Chapter 2
	Introgression between *S. rupestris* and *S. vernalis*	(Comes & Abbott 2001)
	Introgression between *S. vernalis* and *S. glaucus*	(Comes & Abbott 1999) Chapter 3
	Introgression between *S. gallicus* and *S. glaucus*	Chapter 3
	Introgression between *S. vernalis* and *S. leucanthemifolius, S. gallicus* and *S. glaucus* or their common ancestor	Chapter 3
	Introgression between *S. leucanthemifolius* and *S. chrysanthemifolius*	Chapter 3

5.3 The genomic basis for adaptation in *Senecio*

While the proportion of the genome involved in speciation and adaptation is unclear, it is certain that only some subset of it is. But what are the loci of speciation and adaptation themselves? One distinction among speciation and adaptation loci, is that between mutations

in *cis*-regulatory versus protein-coding loci. A recent debate has involved the relative importance of *cis*-regulatory mutations and protein-coding mutations in speciation, and in adaptive evolution more generally. Many researchers have argued that adaptive mutations are overwhelmingly likely to occur in *cis*-regulatory regions (portions of the genome outside coding regions which can alter expression of genes on the same strand without an expressed intermediate; Carroll 2000; Wray 2007; Prud'homme et al. 2007). This view is justified theoretically by the argument that, while a mutation which alters the amino acid sequence of a protein affects that protein wherever it is expressed, the evolution of regulatory regions can affect the expression of a gene in only a certain cell type or under certain environmental conditions (Wray 2007). Thus, regulatory mutations have the potential to exert a smaller number of distinct phenotypic effects (be less pleiotropic); and since most phenotypic alterations to a well-adapted organism are likely to be deleterious, *cis*-regulatory mutations are more likely to be adaptive. However, this view has been highly controversial (Hoekstra & Coyne 2007; Wray 2007; Stern & Orgogozo 2008).

So can results in *Senecio* add to this debate? Previous work identified 64 genes with significant differential expression between *S. aethnensis* and *S. chrysanthemifolius* (Chapman et al. 2013). The study by Chapman (2013) also used several population genetic tests of selection, but these could not differentiate between selection on coding variation and other genomic variation, such as in *cis*-regulatory regions upstream of coding genes. It has also been found that gene expression changes have been important in both the homoploid hybrid speciation of *S. squalidus* and the allopolyploid origin of *S. cambrensis* (Hegarty et al. 2009, 2006). The results of Chapter 4 represent the first study to investigate selection on coding regions in Mediterranean *Senecio* explicitly and indeed are the largest study identifying selection in

coding regions in plants to date. It confirms that some protein-coding genes are under selection in *Senecio* and in nearly all other taxa analysed (with the exception of *Solanum*). In *Senecio,* the 23 protein-coding genes under selection (which are likely to be an underestimate, as discussed in Chapter 4) provide evidence that selection on protein-coding changes has been a feature of their evolution in addition to the expression divergence which was already known in two species of the clade. Making a robust comparison of these two types of genetic change would require identification of a genome-wide sample of substitutions amongst the lineages, and categorising them as residing in either coding or regulatory regions. However, even this approach fails to differentiate between differences fixed neutrally or by selection. Thus, the conclusion which can currently be drawn is that both coding and regulatory changes are likely to have been important in the evolution of Mediterranean *Senecio.* Interestingly, the only study to date to take a genome-wide approach to this question, in a study in sticklebacks (Jones et al. 2012), reached a similar conclusion. Finding that both types of divergence were important, although *cis*-regulatory mutations were more common. The conclusion that both regulatory and coding evolution are key to adaptation and speciation may prove to be correct for most taxa. However, more studies are needed to confirm this conclusion, and to resolve more subtle questions, for example, whether certain classes of gene are more likely to be involved in regulatory or coding evolution than others (Stern and Orgogozo, 2008).

5.4 Conclusions

This thesis has shed light on the prevalence and role of gene flow and selection in *Senecio* as well as in plants more generally. *S. aethnensis* and *S. chrysanthemifolius* appear to be an example of recent ecological speciation with gene flow. The rapid growth of Mount Etna to a

mountain more than 3 km high seems likely to have driven this divergence, and if so the species represent a potent illustration of the power of natural selection to respond to rapid ecological change. However, despite their impressive phenotypic divergence, the species have continued to exchange genes. This propensity for gene flow following divergence is not restricted to these two species. In fact, the results in this thesis suggest that many members of the Mediterranean *Senecio* species-complex have exchanged genes promiscuously following divergence.

Senecio has also experienced a high degree of selection during its divergence, and this has focussed on protein-coding variation in addition to the regulatory divergence that was already known, and is often assumed to predominate. The selection identified in this thesis and previous work is likely to have countered the effects of gene flow in *Senecio*. However, the details of the interaction between these two forces, such as whether gene flow has also played an adaptive role in the divergence of the clade, are yet to be elucidated. Determining the relative role of expression versus coding divergence, and the adaptive or maladaptive nature of gene flow in the species should be a key focus of future research.

The thesis has also made steps towards cementing *Senecio* as one of the most promising systems for the study of speciation and adaptation. This is particularly true of *Senecio aethnensis* and *S. chrysanthemifolius*, which are likely to be a rare example of recent ecological speciation with gene flow and are adapted to a particularly clear ecological cline between their extremely different habitats. The rest of the clade are an outstanding example of the variety of roles that gene flow can play in species diversification. In addition to the advances in understanding of *Senecio,* the studies represented in this thesis provide the first high-throughput sequence data for many of these species, which will undoubtedly prove

useful in future work. Where once interspecific hybridisation was considered an evolutionary oddity, its pre-eminent role in the evolutionary process is now being recognised (Abbott et al. 2013; Seehausen 2004), and the work in this thesis goes some way towards supporting this view.

References

Abbott R et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Abbott R, Brennan A. 2014. Altitudinal gradients, plant hybrid zones and evolutionary novelty. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369:6–9.

Abbott RJ, Ashton PA, Forbes DG. 1992. Introgressive origin of the radiate groundsel, *Senecio vulgaris* L. var. *hibernicus* Syme: *Aat-3* evidence. Heredity. 68: 425–435.

Abbott RJ, James JK, Forbes DG, Comes HP. 2002. Hybrid origin of the Oxford Ragwort , *Senecio squalidus* L : morphological and allozyme differences between *S . squalidus* and *S . rupestris* Waldst. and Kit. Watsonia. 24:17–29.

Alexander JCM. 1979. The Mediterranean species of *Senecio* sections *Senecio* and *Delphinifolius*. Notes from R. Bot. Gard. Edinburgh. 37:387–428.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–10.

Anderson E. 1949. *Introgressive Hybridisation*. John Wiley & Sons, Inc.: New York.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.

Arnold ML et al. 1992. Pollen Dispersal and Interspecific Gene Flow in Louisiana Irises. Heredity. 68:399–404.

Ashton PA, Abbott RJ. 1992. Isozyme evidence and the origin of *Senecio vulgaris* (*Compositae*). Plant Syst. Evol. 179:167–174.

Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM. 2014. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. PLoS Genet. 10:e1004365.

Barker MS et al. 2008. Multiple paleopolyploidizations during the evolution of the *Compositae* reveal parallel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25:2445–55.

Barton NH, Hewitt GM. 1985. Analysis of Hybrid Zones. Annu. Rev. Ecol. Syst. 16:113–148.

Bearhop S et al. 2005. Assortative mating as a mechanism for rapid evolution of a migratory divide. Science. 310:502–504.

De Beni E, Branca S, Coltelli M, Groppelli G, Wijbrans JR. 2011. 40Ar/39Ar isotopic dating of Etna volcanic succession. Ital. J. Geosci. 130:292–305.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B. 57:289-300.

Birky CWJ, Barraclough TG. 2009. Asexual Speciation. In: Lost Sex. The Evolutionary Biology of Parthenogenesis. Van Dijk, P, Martens, K, & Schön, I, editors. Springer pp. 201–216.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric Speciation: Models and Empirical Evidence. Annu. Rev. Ecol. Evol. Syst. 38:459–487.

Borg M, Brownfield L, Twell D. 2009. Male gametophyte development: a molecular perspective. J. Exp. Bot. 60:1465–1478.

Bouckaert RR. 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics. 26:1372–1373.

Boyko AR et al. 2008. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. PLoS Genet. 4:e1000083.

Branca S, Coltelli M, Groppelli G. 2011. Geological evolution of a complex basaltic stratovolcano: Mount Etna, Italy. Ital. J. Geosci. 130:306–317.

Brennan AC, Barker D, Hiscock SJ, Abbott RJ. 2012. Molecular genetic and quantitative trait divergence associated with recent homoploid hybrid speciation: a study of *Senecio squalidus* (*Asteraceae*). Heredity . 108:87–95.

Brennan AC, Bridle JR, Wang A-L, Hiscock SJ, Abbott RJ. 2009. Adaptation and selection in the *Senecio* (*Asteraceae*) hybrid zone on Mount Etna, Sicily. New Phytol. 183:702–17.

Brennan AC, Hiscock SJ, Abbott RJ. 2014. Interspecific crossing and genetic mapping reveal intrinsic genomic incompatibility between two *Senecio* species that form a hybrid zone on Mount Etna, Sicily. Heredity. 113:1–10.

Bridle JR, Vines TH. 2007. Limits to evolution at range margins: when and why does adaptation fail? Trends Ecol. Evol. 22:140–147.

Brown B a et al. 2005. A UV-B-specific signalling component orchestrates plant UV protection. Proc. Natl. Acad. Sci. U. S. A. 102:18225–18230.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 101:577–580.

Chapman MA, Abbott RJ. 2005. The origin of a novel form of *Senecio* (*Asteraceae*) restricted to sand dunes in southern Sicily. New Phytol. 166:1051–62.

Chapman MA, Forbes DG, Abbott RJ. 2005. Pollen competition among two species of *Senecio* (*Asteraceae*) that form a hybrid zone on Mt. Etna, Sicily. Am. J. Bot. 92:730–5.

Chapman MA., Hiscock SJ, Filatov DA. 2016. The genomic bases of morphological divergence and reproductive isolation driven by ecological speciation in *Senecio* (*Asteraceae*). J. Evol. Biol. 29:98-113.

Chapman MA, Abbott RJ. 2010. Introgression of fitness genes across a ploidy barrier. New Phytol. 186:63–71.

Chapman MA, Hiscock SJ, Filatov DA. 2013. Genomic divergence during speciation driven by adaptation to altitude. Mol. Biol. Evol. 30:2553–2567.

Chaw SM, Chang CC, Chen HL, Li WH. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 58:424–441.

Chibalina MV, Filatov DA. 2011. Plant Y Chromosome Degeneration Is Retarded by Haploid Purifying Selection. Curr. Biol. 21:1475–1479.

Coleman M, Abbott RJ. 2003. Possible causes of morphological variation in an endemic Moroccan groundsel (*Senecio leucanthemifolius* var. *casablancae*): Evidence from chloroplast DNA and random amplified polymorphic DNA markers. Mol. Ecol. 12:423–434.

Comes H, Abbott R. 1999. Population genetic structure and gene flow across arid versus mesic environments: A comparative study of two parapatric *Senecio* species from the Near East. Evolution. 53:36–54.

Comes HP, Abbott RJ. 2001. Molecular phylogeography , reticulation , and lineage sorting in Mediterranean *Senecio* sect . *senecio* (A*steraceae*). Evolution 55:1943–1962.

Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674–6.

Coyne JA. 1994. Ernst Mayr and the origin of species. Evolution. 48:19–30.

Coyne JA, Orr HA. 2004. *Speciation*. Sinauer Associates: Sunderland, MA.

Crisp P. 1972. Cytotaxonomic studies in the section *annui* of *Senecio*. PhD thesis. Queen Mary's, University of London.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133–3157.

Cui R et al. 2013. Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. Evolution. 67:2166–2179.

Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray: London.

Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2:318–326.

Defossez E et al. 2011. Do interactions between plant and soil biota change with elevation? A study on *Fagus sylvatica*. Biol. Lett. 7:699–701.

DePristo M a et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Dobzhansky T. 1935. A critique of the species concept in biology. Philos. Sci. 2:344.

Doebeli M, Dieckmann U. 2002. Interim Report: Speciation along environmental gradients. Nature 421: 259-264.

Doorduin L et al. 2011. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris:* SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Res. 18:93–105.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. 102:14338–43.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 10:48.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 107:1–15.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome Res. 24:885–895.

Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. 109:13956–13960.

Eyre-Walker a, Gaut BS. 1997. Correlated rates of synonymous site evolution across plant genomes. Mol. Biol. Evol. 14:455–460.

Feder JL et al. 1994. Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. Proc. Natl. Acad. Sci. U. S. A. 91:7990–7994.

Feder JL, Nosil P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evolution. 64:1729–1747.

Filatov DA. 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. Bioinformatics. 25:3189–90.

Filatov DA, Osborne OG, Papadopulos AST. 2016. Demographic history of speciation in a *Senecio* altitudinal hybrid zone on Mt. Etna. Mol. Ecol. 25:2467–2481.

Filchak KE, Roethele JB, Feder JL. 2000. Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. Nature. 407:739–742.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27:2257–2267.

Fontaine MC et al. 2014. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1–14.

Gompert Z, Alex BC. 2011. Bayesian estimation of genomic clines. Mol. Ecol. 20:2111–2127.

Gossmann TI et al. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol. Biol. Evol. 27:1822–1832.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. 4:658–667.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Haas BJ et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–512.

Hegarty MJ et al. 2009. Extreme changes to gene expression associated with homoploid hybrid speciation. Mol. Ecol. 18:877–889.

Hegarty MJ et al. 2006. Transcriptome Shock after Interspecific Hybridization in *Senecio* Is Ameliorated by Genome Duplication. Curr. Biol. 16:1652–1659.

Hendry AP. 2007. The speed of ecological speciation. Funct. Ecol. 21:455–464.

Hey J. 2006. On the failure of modern species concepts. Trends Ecol. Evol. 21:447–450.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 61:995–1016.

Hopkins R, Rausher MD. 2011. Identification of two genes causing reinforcement in the Texas wildflower *Phlox drummondii*. Nature. 469:411–414.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics. 116:153–159.

Huerta-Sánchez E et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nat. Educ. 512:194–197.

James JK, Abbott RJ. 2005. Recent, allopatric, homoploid hybrid speciation: the origin of *Senecio squalidus* (*Asteraceae*) in the British Isles from a hybrid zone on Mount Etna, Sicily. Evolution. 59:2533-47.

Jones FC et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 484:55–61.

Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol. 29:1125–1139.

Jordan IK, Wolf YI, Koonin E V. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. 4:22.

Kadereit JW, Uribe-Convers S, Westberg E, Comes HP. 2006. Reciprocal hybridization at different times between *Senecio flavus* and *Senecio glaucus* gave rise to two polyploid species in North Africa and South-west Asia. New Phytol. 169:431–441.

Kane NC, Barker MS, Zhan SH, Rieseberg LH. 2011. Molecular evolution across the *Asteraceae*: micro- and macroevolutionary processes. Mol. Biol. Evol. 28:3225–35.

Kapralov MV, Votintseva AA, Filatov DA. 2013. Molecular adaptation during a rapid adaptive radiation. Mol. Biol. Evol. 30:1051–1059.

Kim M et al. 2008. Regulatory genes control a key morphological and ecological trait transferred between species. Science. 322:1116–9.

Kimura M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press: Cambridge, UK.

Kleindorfer S et al. 2014. Species collapse via hybridization in Darwin's tree finches. Am. Nat. 183:325–41.

Kondrashov AS, Kondrashov FA. 1999. Interactions among quantitative traits in the course of sympatric speciation. Nature. 400:351–354.

Koonin E V. 2011. Are there laws of genome evolution? PLoS Comput. Biol. 7:e1002173.

Koonin E V, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr. Opin. Biotechnol. 17:481–7.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 15:356.

Körner C. 2007. The use of "altitude" in ecological research. Trends Ecol. Evol. 22:569–574.

Kosiol C et al. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genet. 4:e1000144.

Kozarewa I et al. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Methods. 6:291–295.

Kunte, K., and D. Agashe. 2015. Preface - Special Section: Evolutionary Biology. Current Science. 108:1835-1873

Kutschera VE et al. 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol. Biol. Evol. 31:2004–2017.

Larracuente AM et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.

Lawton-Rauh A, Robichaux RH, Purugganan MD. 2007. Diversity and divergence patterns in regulatory genes suggest differential gene flow in recently derived species of the Hawaiian silversword alliance adaptive radiation (*Asteraceae*). Mol. Ecol. 16:3995–4013.

Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: A simulation study. Syst. Biol. 63:17–30.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

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

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. Evolution. 62:2984–2994.

Lowe AJ, Abbott RJ. 1996. Origins of the new allopolyploid species *Senecio cambrensis* (*Asteraceae*) and its relationship to the Canary Islands endemic *Senecio teneriffae.* Am. J. Bot. 83:1365–1372.

Lowe AJ, Abbott RJ. 2004. Reproductive isolation of a new hybrid species, *Senecio eboracensis* Abbott & Lowe (*Asteraceae*). Heredity. 92:386–395.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat. Rev. Genet. 4:981–994.

Malhis N, Jones SJM. 2010. High quality SNP calling using Illumina data at shallow coverage. Bioinformatics. 26:1029–1035.

Mallick S et al. 2009. The difficulty of avoiding false positives in genome scans for natural selection populations. Genome Res. 19:922–933.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17:10–12.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32:244–257.

Mayr E. 1963. *Animal Species and Evolution*. Belknap: Cambridge, MA.

Mayr E. 1957. Species concepts and definitions. In: The Species Problem. Mayr, E, editor. AAAS: Washington, DC. pp. 371–388.

Mayr E. 1942. *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press: Cambridge, MA.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 351:652–654.

Mirarab S et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–i548.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. 8:122–8.

Morjan CL, Rieseberg LH. 2004. How species evolve collectively: Implications of gene flow and selection for the spread of advantageous alleles. Mol. Ecol. 13:1341–1356.

Muir G, Dixon CJ, Harper AL, Filatov DA. 2012. Dynamics of drift, gene flow, and selection during speciation in *Silene*. Evolution. 66:1447-1458.

Muir G, Osborne OG, Sarasa J, Hiscock SJ, Filatov DA. 2013. Recent ecological selection on regulatory divergence is shaping clinal variation in *Senecio* on Mount Etna. Evolution. 67:3032–3042.

Neafsey DE et al. 2010. Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. Genome Res. 20:938–946.

Nevado B, Fazalova V, Backeljau T, Hanssens M, Verheyen E. 2011. Repeated unidirectional introgression of nuclear and mitochondrial dna between four congeneric Tanganyikan cichlids. Mol. Biol. Evol. 28:2253–2267.

Nielsen R et al. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the *HIV-1* envelope gene. Genetics. 148:929–936.

Noor MAF, Bennett SM. 2010. Islands of Speciation or Mirages in the Desert? Examining the Role of Restricted Recombination in Maintaining Species. Heredity. 103:439–444.

Nosil P, Harmon LJ, Seehausen O. 2009. Ecological explanations for (incomplete) speciation. Trends Ecol. Evol. 24:145–156.

Nosil P, Vines TH, Funk DJ. 2005. Perspective: Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution. 59:705–719.

Oberprieler C, Heine G, Bässler C. 2015. Can divergent selection save the rare *Senecio hercynicus* from genetic swamping by its spreading congener *S. ovatus* (*Compositae, Senecioneae*)? Flora - Morphol. Distrib. Funct. Ecol. Plants. 210:47–59.

Oh S, Zhang H, Ludwig P, Nocker S Van. 2004. A mechanism related to the yeast transcriptional regulator Paf1c is required for expression of the *Arabidopsis* FLC/MAF MADS Box gene family. Plant Cell. 16:2940–2953.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag: New York.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:263–286.

Osborne OG, Batstone TE, Hiscock SJ, Filatov DA. 2013. Rapid speciation with gene flow following the formation of Mt. Etna. Genome Biol. Evol. 5:1704–1715.

Osborne OG, Chapman MA, Nevado B, Filatov DA. 2016. Maintenance of species boundaries despite ongoing gene flow in ragworts. Genome Biol. Evol. 8: 1038-47

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5(5):568-583.

Panero JL, Funk VA. 2008. The value of sampling anomalous taxa in phylogenetic studies: Major clades of the *Asteraceae* revealed. Mol. Phylogenet. Evol. 47:757–782.

Papadopulos AST et al. 2013. A comparative analysis of the mechanisms underlying speciation on Lord Howe Island. J. Evol. Biol. 26:733–745.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.

Pardo-Diaz C et al. 2012. Adaptive introgression across species boundaries in Heliconius butterflies. PLoS Genet. 8:e1002752.

Parker J et al. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. Nature. 502:1–9.

Pelser PB et al. 2012. The genetic ghost of an invasion past: Colonization and extinction revealed by historical hybridization in *Senecio*. Mol. Ecol. 21:369–387.

Pelser PB, Nordenstam B, Kadereit JW, Watson LE. 2007. An ITS phylogeny of tribe *Senecioneae* (*Asteraceae*) and a new delimitation of Senecio L. Taxon. 56:1077–1104.

Pie MR. 2006. The influence of phylogenetic uncertainty on the detection of positive Darwinian selection. Mol. Biol. Evol. 23:2274–2278.

Pinho C, Hey J. 2010. Divergence with Gene Flow: Models and Data. Annu. Rev. Ecol. Evol. Syst. 41:215–230.

Poelstra JW et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 344:1410–1414.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc. Natl. Acad. Sci. U. S. A. 104:8605–8612.

Renaut S et al. 2012. Genome-wide patterns of divergence during speciation: the lake whitefish case study. Philos. Trans. R. Soc. B Biol. Sci. 367:354–363.

Rhymer JM, Simberloff D. 1996. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27:83–109.

Rieseberg LH. 2006. Hybrid Speciation in Wild Sunflowers. Ann. Missouri Bot. Gard. 93:34–48.

Roda F et al. 2013. Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. Mol. Ecol. 22:2941–2952.

Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol. Evol. 24:192–200.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol. 4:1899–1904.

Ross RIC. 2010. Local adaptation and adaptive divergence in a hybrid species complex in *Senecio*. PhD thesis. University of Oxford.

Ross RIC, Ågren JA, Pannell JR. 2012. Exogenous selection shapes germination behaviour and seedling traits of populations at different altitudes in a *Senecio* hybrid zone. Ann. Bot. 110:1439–1447.

Rundle H, Nosil P. 2005. Ecological speciation. Ecol. Lett. 336–352.

Salathé M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol. Biol. Evol. 23:721–722.

Sambatti JBM, Strasburg JL, Ortiz-Barrientos D, Baack EJ, Rieseberg LH. 2012. Reconciling extremely strong barriers with high levels of gene exchange in annual sunflowers. Evolution. 66:1459–1473.

Savolainen V et al. 2006. Sympatric speciation in palms on an oceanic island. Nature. 441:9–12.

Schliep KP. 2011. Phangorn: Phylogenetic analysis in R. Bioinformatics. 27:592–593.

Schliewen UK, Tautz D, Pääbo S. 1994. Sympatric speciation suggested by monophyly of crater lake cichlids. Nature. 368:629–632.

Seehausen O. 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19:198–207.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Shimodaira H, Hasegawa M. 1999. Letter to the Editor: Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Silvertown J, Servaes C, Biss P, Macleod D. 2005. Reinforcement of reproductive isolation between adjacent populations in the Park Grass Experiment. Heredity. 95:198–205.

Sironi M, Cagliani R, Forni D, Clerici M. 2015. Evolutionary insights into host–pathogen interactions from mammalian sequence data. Nat. Rev. Genet. 16:224–236.

Slotte T et al. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. Genome Biol. Evol. 3:1210–9.

Smadja CM, Butlin RK. 2011. A framework for comparing processes of speciation in the presence of gene flow. Mol. Ecol. 20:5123–5140.

Soria-Carrasco V et al. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. Science. 344:630–634.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Stern DL. 2013. The genetic causes of convergent evolution. Nat. Rev. Genet. 14:751–64.

Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evolution? Evolution. 62:2155–2177.

Strasburg JL et al. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol. Biol. Evol. 28:1569–1580.

Strasburg JL, Rieseberg LH. 2008. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris* - Large effective population sizes and rates of long-term gene flow. Evolution. 62:1936–1950.

Sun Y et al. 2012. Large-scale introgression shapes the evolution of the mating-type chromosomes of the filamentous ascomycete *Neurospora tetrasperma*. PLoS Genet. 8:e1002820.

Surget-Groba Y, Montoya-Burgos JI. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 20:1432–1440.

Swarbreck D et al. 2008. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36:1009–14.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 135:599–607.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Tamura K et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

The Heliconius Genome Consortium et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 487:94–98.

Turner TL, Hahn MW. 2010. PERSPECTIVE: Genomic islands of speciation or genomic islands and speciation? Mol. Ecol. 19:848–850.

Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. Heredity. 108:179–189.

Warren DL, Glor RE, Turelli M. 2010. ENMTools: A toolbox for comparative studies of environmental niche models. Ecography. 33:607–611.

Webb WC, Marzluff JM, Omland KE. 2011. Random interbreeding between cryptic lineages of the common raven: Evidence for speciation in reverse. Mol. Ecol. 20:2390–2402.

Whitney KD, Ahern JR, Campbell LG, Albert LP, King MS. 2010. Patterns of hybridization in plants. Perspect. Plant Ecol. Evol. Syst. 12:175–182.

Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84:9054–9058.

Wray G a. 2007. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. 8:206–216.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 97–159.

Wu C. 2001. The genic view of the process of speciation. 14:851–865.

Wyatt R, Odrzykoski IJ, Stoneburner a, Bass HW, Galau G a. 1988. Allopolyploidy in bryophytes: Multiple origins of *Plagiomnium medium*. Proc. Natl. Acad. Sci. U. S. A. 85:5601–4.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol. Biol. Evol. 28:2359–2369.

Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. Genome Biol. Evol. 2:200–11.

Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics. 162:1811–1823.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–91.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–17.

Yang Z, Nielsen R, Goldman N, Krabbe Pederson A-M. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Mol. Biol. Evol. 19:49–57.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–18.

Zardi GI et al. 2011. Adaptive traits are maintained on steep selective gradients despite gene flow and hybridization in the intertidal zone. PLoS One. 6:e0019402.

Zhai W, Nielsen R, Goldman N, Yang Z. 2012. Looking for Darwin in genomic sequences-validity and success of statistical methods. Mol. Biol. Evol. 29:2889–2893.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–9.