Speciation and Adaptation of *Senecio*



Lok Yin Edgar Wong

Linacre College, Oxford

A thesis submitted for the degree of Doctor of Philosophy, Michaelmas term, 2020

Under the supervision of Professor Dmitry A. Filatov and Professor Simon, J. Hiscock

Department of Plant Sciences

Acknowledgements

The journey of my D.Phil. would have been very difficult and challenging without so many people's help and support.

First and foremost, I would like to thank my supervisors, Dmitry and Simon, for sharing with me tremendous amount of knowledge and advice, intellectually challenging me all the time, pushing and guiding me through the whole process. Bruno deserves a special shout-out, as he is essentially my unofficial third supervisor who has given me so much technical support and guidance. A big thank you to my friends and colleagues in the department: Marc, James, Mahdi, Ying-Qi who are in the research group; Pablo, Tom Carruthers, Tom Wells in the office; Pauline and Helen in the lab; Tim and Eric in the greenhouse; Gem as our caring graduate administrator; and many other researchers in Oxford or somewhere else, graduate students and all other support staff in the Department of Plant Sciences.

I have also received so much support and love from my close friends in the UK: Heather, Gareth, Henry, Aon, Tiffany, Brandi, Ari, Oli, Jamie, etc; and Hong Kong: David, Steven, Jaime, Ivan, Penny, Helen, Rachel, etc. They always make me feel that I am not alone in this journey. In addition to all my friends, I ought to thank my Biology teacher of five years in secondary school, Miss Wong, who gave me the very first taste of being a biologist and many opportunities to explore my scientific potentials in my teenage years. My BSc and MSc dissertation supervisors, Richard Saunders and Alex Twyford, were also very important role models in my scientific journey as they have always been extremely encouraging and intellectually inspiring when I worked in their groups.

Most importantly, I want to express my heartfelt gratitude to my parents, sister and my extended family for their love and support both financially and emotionally in the last four years. Without them, I would not be able to achieve this.

Contents

1. Thesis Abstract 1
2. General Introduction 2
   - 2.1 What is speciation? 2
   - 2.2 Role of local adaptation and divergent selection in reproductive isolation and speciation 4
   - 2.3 Role of hybrid zones and gene flow in speciation 5
   - 2.4 Approaches in studying adaptation and speciation 8
   - 2.5 The study system – Etnean *Senecio* and the homoploid hybrid species in the UK 9
   - 2.6 Aims of research 10
3. Strong divergent selection at multiple loci in two closely related species of ragworts adapted to high and low elevations on Mount Etna 14
   - 3.1 Preamble 15
   - 3.2 Abstract 15
   - 3.3 Introduction 16
   - 3.4 Materials and Methods 20
     - 3.4.1 Sampling 20
     - 3.4.2 nextRAD sequencing 21
     - 3.4.3 Analysis of raw reads, genetic diversity and structure in nextRAD data set 21
     - 3.4.4 Demographic modelling using nextRAD data 22
     - 3.4.5 Measurement of leaf phenotype 26
     - 3.4.6 Identification of outlier loci in the nextRAD data set 26
     - 3.4.7 Cline fitting 27
     - 3.4.8 Strength of selection 29
     - 3.4.9 Effect of selection on wider genomic regions 29
   - 3.5 Results 30
     - 3.5.1 Genetic diversity and structure of localities inferred from nextRAD data set 30
     - 3.5.2 Demographic history of *Senecio* on Mount Etna 34
     - 3.5.3 Cline in leaf dissection 38
     - 3.5.4 Identification of outlier nextRAD markers and analysis of their clines 38
     - 3.5.5 Potential association between outlier markers and functional genes 41
     - 3.5.6 Strength of selection 51
     - 3.5.7 How wide are the genomic regions affected by divergent selection? 52
   - 3.6 Discussion 53
     - 3.6.1 Demographic history of speciation in Etnean *Senecio* 53
     - 3.6.2 Variable coincidence and concordance among phenotypic and genotypic clines 54

3.6.3 Strong selection against hybrids on Mount Etna 55
3.6.4 Intrinsic and extrinsic causes in the divergence of Etnean *Senecio* 57
3.6.5 Maintenance of divergence despite gene flow 58
4. Genome scans reveal vast number of regions harbouring putative selective sweeps and genes related to adaptation in high- and low-elevation *Senecio* species on Mount Etna 60
4.1 Abstract 61
4.2 Introduction 62
4.3 Materials and Methods 66
4.3.1 Integrating the linkage map with genomic sequences 66
4.3.2 Analyses of polymorphisms in nextRAD markers and differentially expressed genes 66
4.3.3 Filtering of raw reads, SNP calling and sliding window analyses in the *S. aethnensis* and *S. chrysanthemifolius* genome sequences 68
4.3.4 Genomic scans for signals of selective sweeps and subsequent analyses of selection strength and age 69
4.3.5 Gene Ontology enrichment analysis 70
4.4 Results 71
4.4.1 Correlation between recombination rate and DNA polymorphism 71
4.4.2 Sliding windows of polymorphisms and selective sweep scans in genome sequences 77
4.4.3 Strength and age of chosen selective sweeps 78
4.4.4 Genes harboured in regions with strong putative selective sweeps 79
4.4.5 Gene Ontology enrichment analyses 80
4.5 Discussion 116
4.5.1 Correlation between recombination rate and DNA polymorphism 116
4.5.2 Signals of putative selective sweeps and patterns of polymorphism 117
4.5.3 False-positives in genome scans 119
4.5.4 Strength and age of selective sweeps 119
4.5.5 Genes in genomic regions with putative sweeps and their implications for adaptation 120
4.5.6 Insights into selection and adaptation at high and low elevations 122
5. Multiple synthetic crosses confirmed hybrid breakdown in Etnean *Senecio* and significant differences from the natural hybrid, *Senecio squalidus* – insights from phenotypic traits 125
5.1 Abstract 126
5.2 Introduction 127
5.3 Materials and Methods 131
5.3.1 Crossing design 131
5.3.2 Seed germination and plant rearing conditions 131

5.3.3 *Senecio squalidus* samples 132
5.3.4 Leaf traits 133
5.3.5 Other phenotypic characterisation in F_2 individuals 133
5.3.6 Statistical analyses 133
5.4 Results 135
5.4.1 Survival and signs of hybrid breakdown 135
5.4.2 Differences in phenotypic characteristics among F_2 groups 138
5.4.3 Differences in phenotypic characteristics between the two F_2 families derived from a reciprocal F_0 cross 144
5.4.4 Differences in leaf traits between synthetic F_2 hybrids and the hybrid species *S. squalidus* 144
5.4.5 Differences in leaf traits between synthetic F_2 hybrids and *S. squalidus* - Exeter samples grown in the greenhouse 148
5.4.6 Differences in leaf traits between wild and greenhouse-grown *S. squalidus* under identical genetic background 148
5.5 Discussion 149
5.5.1 Signs of hybrid breakdown 149
5.5.2 Leaf architecture, floral and vegetative characteristics in F_2 hybrids 150
5.5.3 Nucleocytoplasmic effect on F_2 hybrids 151
5.5.4 Comparing natural and synthetic hybrids 152
5.5.5 Effect of synthetic environments on phenotypes 152
5.5.6 Implications for hybrid reproductivity and viability on Mount Etna 153
6. General Discussion 155
6.1 Demographic history and divergence in Etnean Senecio 155
6.2 Selective mechanisms in Etnean *Senecio* 157
6.3 Hybrids, hybridisation and maintenance of the hybrid zone 158
6.4 Conclusion - Etnean *Senecio* as a model to study speciation with gene flow 161
7. References 162

1. Thesis Abstract

Speciation and Adaptation of *Senecio*

Lok Yin Edgar Wong, Linacre College, Oxford. A thesis submitted for the degree of Doctor of Philosophy, Michaelmas term, 2020.

Speciation was thought to be initiated by geographical isolation only. However, more and more research has shown that there are other drivers of speciation, such as divergent selection, and that speciation could happen in the face of gene flow with mechanisms to maintain reproductive isolation. This thesis presents similar findings on a case of recent speciation with gene flow found on Mount Etna (Sicily), where closely related high- and low-elevation adapted species, *Senecio aethnensis* and *S. chrysanthemifolius* live and hybridise. I used different methods to examine the processes during their speciation and the maintenance of their reproductive isolation. Their homoploid hybrid species, *S. squalidus*, was also studied to explore various aspects of hybrid speciation. I started by exploring the demographic history of the two species, and revealed several crucial features during their speciation, including heterogenous gene flow, bidirectional gene flow and historical gene flow. Although a secondary contact scenario and a gene flow since divergence scenario could not be determined, the rest of the thesis focused on the evolutionary processes that happened during speciation. Cline analyses and genome scans revealed the presence of divergent selection and positive selection widespread across all linkage groups and that selective agents had various strengths and ages, supporting the multifarious selection hypothesis. These two analyses also provided a list of candidate genes that are related to adaptation to different elevations on Mount Etna, such as those regarding metal ions defence and UV response. In the final chapter of the thesis, I investigated hybrid breakdown using multiple F_2 families. Not only was hybrid breakdown confirmed, new breakdown traits were discovered. I also uncovered phenotypic evidence of nucleocytoplasmic incompatibility in the system, and showed that the level of hybrid breakdown likely depended on the level of incompatibility between different *S. aethnensis* and *S. chrysanthemifolius* pairs. Last but not least, the hybrid species *ex situ*, *S. squalidus*, exhibited significant phenotypic differences from artificial F_2 hybrids, hinting morphological novelties during its evolution and establishment in the UK. Overall, this thesis presents research on both the intrinsic and extrinsic mechanisms during the speciation of Etnean *Senecio.* At the same time, it illustrates the value of these young non-model systems in studying speciation and demonstrates that evolution of new traits for adaptation can happen very quickly.

2. General Introduction

Ever since Darwin wondered about the origin of species in his book *On the Origin of Species* (and probably before that), biologists have not stopped looking for answers for how species originate. The more research is done, the more complicated the question becomes. Perhaps this is because biology is a science of exceptions – there is no one rule that applies to all. Cases that are similar in many aspects often have other features that are distinctive to each of them. This is especially true for speciation studies – from the start of divergence, to accumulation of divergence and adaptation to their respective environments, to finally being reproductive isolated and stable, millions of processes could have happened. There could be multiple extinctions of diverging lineages, hybridisation with other closely-related taxa, geographical events that separate or bring together different populations, hundreds and thousands of beneficial and deleterious mutations, selective agents at different stages of the speciation, and so on. It is crucial to carefully scrutinise each process involved in order to get a real picture of speciation in each system and add novel, exciting information to the body of knowledge in the field.

In this chapter, I will outline some of the things evolutionary biologists have learnt about speciation over the years and methods that can be applied. I will also introduce the system known to be a case of speciation with gene flow and end by describing how I use this system to answer outstanding questions in the field of speciation and adaptation.

2.1 What is speciation?

Speciation is the process in which new species are formed. To understand speciation, one must first understand what a species is. There are many definitions of a species: the 'biological species concept' (BSC) defines a species to be reproductively isolated from others and one that would not produce fertile offspring with other species (Mayr, 1942); conventional taxonomists classify species from a morphological point of view based on their morphological distinctions; whereas others define species from an ecological perspective and see species as the result of ecologically-based divergent selection (Rundle & Nosil, 2005). Contrary to the BSC, more and more genetic studies have revealed

that different, but usually closely-related, 'species' could exchange genes while maintaining substantial differences. Some of these species are considered to be distinct identities if they possess enough differences with regard to ecology, morphology and genetics, even if they are not completely reproductively isolated. This is true for species complexes, in which several cryptic species share both features of 'identical' and 'different' species [such as the *Mimulus guttatus* (e.g. Fishman, Kelly, Morgan, & Willis, 2001) and *Heliconius* (e.g. Hines *et al.*, 2011) species complexes]. There are also species in the 'speciation continuum', in which they have accumulated substantial amount of differences but their reproductive isolation is not yet complete. In this thesis, two of the species in focus fall in this category. They have accumulated certain ecological, morphological and genetic disparities, yet they still produce offspring (with reduced fertility).

Perhaps one of the reasons why defining a species is so challenging is because there are many ways to achieve speciation as each species would experience their unique combination of extrinsic, intrinsic and demographic processes. Extrinsically, populations, and eventually species, can accumulate differences simply by geographical separation (such as allopatric or parapatric speciation), for example seeds being carried to a novel environment such as an island; speciation can also happen in the same geographical location (sympatric speciation) if different groups are exposed to different environmental conditions such as different substrates. Darwin's theory of natural selection (Darwin, 1859) echoes this, as he suggested divergent selection could lead to differentiation between groups of organisms adapted to each of their local habitat. Intrinsically, species could undergo polyploidisation through hybridisation or whole genome duplications, or simply lots of mutations, to create new genetic identities which might turn into new species under suitable circumstances. Demographically, different events in the population history of species would leave different signals to be detected genetically based on the complex interplay among factors such as the amount of gene flow. For example, markers not affected by gene flow between two species would have similar divergence times, whereas markers that experience gene flow would have reduced divergence compared to the rest of the genome. Of course, speciation is not always a success. It is the 'survival of the fittest' (Spencer, 1864) that leaves the species that we see through fossil and herbarium records and in nature that allow evolutionary biologists to study in order to test different theories regarding speciation and evolution.

2.2 Role of local adaptation and divergent selection in reproductive isolation and speciation

No matter how scientists define species, there is consensus that a certain amount of reproductive isolation should be present in order for species to be considered different. Reproductive isolation exists in many forms: including pre-mating pre-zygotic isolation [habitat or temporal isolation, selection against immigrants, sexual isolation (pollinators and sexual selection against hybrids)], post-mating pre-zygotic isolation, and post-zygotic isolation (intrinsic or ecology-dependent) (Rundle & Nosil, 2005). It is important to note that reproductive isolation is not the cause of diversification, but rather a by-product by divergent selection and genetic drift (Rieseberg & Willis, 2007). Divergent selection causes populations to evolve local genotypes that perform well in their local environment but not in foreign ones (Kawecki & Ebert, 2004). It can also further promote genetic divergence by acting on loci that are linked to those under direct selective pressure, hence exerting an indirect yet widespread impact on genome-wide neutral divergence (Nosil, Funk, & Ortiz-Barrientos, 2009). Similar to reproductive isolation, divergent selection exists in many forms extrinsically and intrinsically. Extrinsic selection is relatively easy to detect, since it would manifest as differences in phenotypic traits of organisms, which could then be linked to the corresponding differences in ecology and environmental conditions. This allows candidate traits and genes to be identified for in-depth studies. Various differences in environmental conditions have been shown to exert divergent selection leading to divergence, such as latitude (Riihimäki & Savolainen, 2004), altitude (Chapman, Hiscock, & Filatov, 2013; Montejo-Kovacevich *et al.*, 2019), and areas of tidal zones (Le Pennec *et al.*, 2017; Westram *et al.*, 2018). Compared to extrinsic selection, intrinsic selection, the genetic incompatibilities leading to reduced fitness of hybrids (e.g. due to developmental defects), is more challenging to detect since there are no obvious candidate genes for functional analyses (Rieseberg & Willis, 2007). In such case, candidates (including extrinsic ones) could be located through genome-wide scans, an approach known as 'reverse ecology' (Li, Costello, Holloway, & Hahn, 2008), which is employed in this thesis. Although intrinsic selection requires more effort to study, model system often provides some insights and references as to why intrinsic incompatibilities arise (e.g. *Arabidopsis* species, reviewed in Vaid & Laitinen, 2019). Research on different species and accessions of *Arabidopsis* has shown that reduced

fitness of hybrids could be caused by interactions between gene variants (e.g. necrosis) (e.g. Bomblies & Weigel, 2007), changes in DNA methylation patterns in hybrids (e.g. Groszmann *et al*., 2011; Shen *et al*., 2012), and nucleocytoplasmic incompatibilities (e.g. Johnson, 2010; Bohra *et al.*, 2016). *Chapter 5* of this thesis aimed to look for manifestation of any of these phenomena in artificial hybrids created in the greenhouse.

Selection to achieve local adaption could occur on multiple targets and with varying strengths. The questions that arise are how many of them there are and how much each of them contributes to reproductive isolation and eventually speciation – the classic debate of whether adaptive divergence is a result of a few loci of large effects (Gavrilets & Losos, 2009) or many loci of small effects (Fisher, 1930). This is in line with the stronger/multifarious selection hypothesis, which proposes that divergence can be caused by strong selection on one trait or multifarious selection on multiple traits (Nosil, Harmon, & Seehausen, 2009). After divergent selection has caused substantial amount of restrictions in gene flow, reproductive isolation will then be reinforced through independent genetic drift, and directional and/or balancing selection in regions of divergence hitchhiking in the later stage of speciation (Via, 2009). Positive selection helps to increase divergence as it drives beneficial adaptation genes (or 'speciation genes') and those in tight linkage to high frequency while balancing selection removes deleterious copies. As gene flow reduces, genetic drift will start to further drive divergence in regions that are not directly under selection, through the accumulation of Bateson–Dobzhansky–Muller (BDM) incompatibilities. BDM incompatibilities cause negative epistatic interactions between genes with different population history (as a consequence of isolation) and are thought to be one of the most common causes of intrinsic post-zygotic reproductive isolation (Seehausen *et al.*, 2014).

2.3 Role of hybrid zones and gene flow in speciation

It is not uncommon that species that are reproductively isolated can still exchange genes and produce hybrids. Hybridization has been studied in many plants species and has been considered an important factor in evolution and speciation by botanists such as G. Ledyard Stebbins (e.g. Anderson & Stebbins, 1954; Stebbins, 1950). However, it has traditionally been disregarded by zoologists as it

was seen as rather unnatural and uncommon (e.g. Mayr, 1999; and other zoologists heavily influenced by his views: Mallet, 2005). In the past, it was also considered unlikely for speciation to happened in the presence of gene flow because recombination would break down linkage disequilibrium between differentiated genes and stop divergence from building up (Bolnick & Fitzpatrick, 2007). However, as more and more cases of hybridisation have been studied closely in both animals and plants, its role in speciation is now widely accepted. Even though gene flow homogenises the genomes of the species that are exchanging genes, there are mechanisms in which the impeding effect of recombination can be overcome. For instance, regions under divergent selection would have low effective recombination rate (Via & West, 2008); tight physical linkage (Butlin, 2005) and pleiotropy (Kirkpatrick & Barton, 1997; Kirkpatrick & Ravigné, 2002) could also assist in maintaining divergence in the genes and genomic regions involved. Sometimes, gene flow can even facilitate speciation by increasing genetic diversity that selection can act on (Mallet, 2005; Smadja & Butlin, 2011). In fact, speciation with gene flow has been widely reported in multiple systems, such as butterflies (*Heliconius*: e.g. Mallet, Beltrán, Neukirchen, & Linares, 2007; Nadeau *et al.*, 2012), equids (*Equus*: e.g. Jónsson *et al.*, 2014), sticklebacks (*Gasterosteus*: e.g. Vines *et al.*, 2016), monkeyflowers (*Mimulus*: e.g. Ferris, Barnett, Blackman, & Willis, 2017; Fishman, Stathos, Beardsley, Williams, & Hill, 2013), and so on.

As hybrids are the product of gene flow between diversified population and species, hybrids and hybrid zones offer valuable opportunities for evolutionary biologists to study the processes that are involved in reproductive isolation and species divergence. Hybrid zones are maintained by the balance between dispersal and selection against hybrids (Barton & Hewitt, 1985). One of the methods to utilise hybrid zones is cline analyses, which involved examining the transition in allele frequencies or phenotypic traits that are associated with reproductive isolation along an environmental gradient (such as latitude, altitude, tidal zones, and so on) between diverged groups and along hybrid zones. Classic cline theory has provided a framework for studying the dynamics of reproductive isolation with the presence of gene flow (Abbott *et al.*, 2013). Barton & Hewitt (1985) pointed out that associations among lots of loci involved in barriers to gene flow, spread across the genome and likely involved in many traits are necessary to enable significant divergence in the genome on each side of the consensus cline. Through analysing different parameters of clines of different traits and markers, those under intrinsic and extrinsic selection and those involved in pre- or post-mating isolation, could

be identified as they would result in steep transitions between the two ends of the environmental gradient. These clines' shapes could also coincide with those of environmental variables, suggesting links between matching clines (although unrelated environmental variables can have similar transitions, and loci that are linked to those under extrinsic selection would show similar transitions as well). Apart from cline analysis, hybrid zones can be used in multiple other ways, such as in quantitative trait analyses. In non-model systems in which genomic resources are scarce, the genome of the hybrid can be used as a reference genome for its progenitors. In some cases, it can also be used to locate genomic structural differences, such as chromosomal rearrangements, between its progenitors (e.g. Fishman *et al.*, 2013). Through studying how hybrid zones are maintained, one will start to get a glimpse of the extrinsic and intrinsic reproductive isolation involved in species divergence and the factors preventing the fusion of species. In *Chapter 3* of this thesis, cline analysis will be employed and discussed further.

Not only can gene flow lead to hybrids, it can also lead to formation of new species through hybrid speciation, where hybrids overcome intrinsic incompatibilities stemmed from their progenitors and become reproductively isolated. The hybrid swarm theory is in favour of hybridisation's role in adaptive radiations and creation of new species by increasing genetic variations in the pool of hybrids for adapting to different niches (peaks) in an adaptive landscape (Seehausen, 2004). Hybridisation can also facilitate divergence by releasing diverging populations from constraints caused by genetic correlations (Seehausen *et al.*, 2014). Sometimes, hybrid species would occupy a novel niche and outperform their progenitors there. An excellent example of this is the sunflower *Helianthus paradoxus*, which has a high salt tolerance and occupies salt marshes even though its progenitor species are both sensitive to salt (Lexer, Welch, Durphy, & Rieseberg, 2003). There are four potential obstacles hybrids ought to overcome before becoming a 'species', including ploidy level, chromosomal rearrangement, genetic incompatibilities and interactions between nuclear genome and endosymbionts (Wolf, Lindell, & Backström, 2010). Regarding ploidy level, hybrids might undergo allopolyploidisation (the fusion of unreduced gametes) to overcome the different ploidy levels of the two genomes of progenitors. Allopolyploidisation is well-documented in plants (e.g. Estep *et al.,* 2014; Lee, Mummenhoff, & Bowman, 2002; Joly, Heenan, & Lockhart, 2009; Soltis & Soltis, 2009; Triplett, Clark, Fisher, & Wen, 2014) but not uncommon in other organisms (e.g. Hotz, Semlitsch, Gutmann,

Guex, & Beerli, 1999; Lampert & Schartl, 2008; Matos, Machado, Schartl, & Coelho, 2015; Stöck *et al.*, 2002). In the case of homoploid speciation (speciation without change in ploidy level), all four pose a huge challenge if a novel hybrid genome is to become genetically stable and reproductively isolated from its parents (Seehausen, 2004). However, experiments and computer simulations have shown that genome stabilisation in homoploid hybrid species can be achieved in a relatively short period of fewer than 25 generations (Ungerer, Baird, Pan, & Rieseberg, 1998).

2.4 Approaches in studying adaptation and speciation

Schluter (2009) outlined top-down and bottom-up approaches in studying the mechanisms underlying speciation and selection. Top-down approaches are informed by phenotypes known to be involved in adaptation (Wolf *et al.*, 2010). These include identifying phenotypic traits that are under divergent selection, traits that are linked to reproductive isolation, and genes underlying these traits and reproductive isolation (Schluter, 2009). These are useful if extensive phenotypic data and genomic resources are available, such as those in model organisms.

With the advancement in next-generation sequencing technologies and their cost-effectiveness, bottom-up approaches are common especially for non-model systems in which little is known about both their phenotypes and genetics. Bottom-up approaches are evolutionary genomics studies that are carried out without *a priori* knowledge on phenotypes (Wolf *et al.*, 2010), including identifying alternative alleles at loci under divergent selection and phenotypes that show differences in fitness, testing whether candidate genes display signatures of positive selection, and genetic mapping of reproductive isolation between closely related species (Schluter, 2009). Cline analysis falls in this category. Genetic mapping studies involves genome scans that look for signatures of positive selection or 'islands of speciation' (Turner, Hahn, & Nuzhdin, 2005), which are regions experiencing restricted gene flow due to differentiation between species. Genes and linked variations within those speciation islands can then be identified for further analyses and experiments to test their effects on adaptation and fitness between the diverged species.

In this thesis, bottom-up approaches such as cline analysis and genome scans will be used; a phenotypic study aiming to explore the conflict resulting from mixing genomes of two species will also be carried out.

2.5 The study system – Etnean *Senecio* and the homoploid hybrid species in the UK

In this thesis, I focus on *Senecio aethnensis* Jan ex. DC., *Senecio chrysanthemifolius* Poir., their natural hybrids on Mount Etna, their synthetic hybrid progenies produced in the greenhouse, and the hybrid species between the two in the UK (*Senecio squalidus*).

Typical *S. aethnensis* occurs above 2,000 m whereas typical *S. chrysanthemifolius* occurs below 1,000 m (Brennan *et al*., 2009; Muir *et al*., 2013*).* The two species maintain a hybrid zone at their intermediate elevations (Chapman, Forbes, & Abbott, 2005; James & Abbott, 2005), where they exhibit a range of intermediate phenotypes between the two species (Brennan, Bridle, Wang, Hiscock, & Abbott, 2009; James & Abbott, 2005). *S. aethnensis* and *S. chrysanthemifolius*' habitats show contrasting environmental conditions, both elevation- and ecology-related, such as temperature, solar radiation, and water availability (James & Abbott, 2005; Körner, 2007; Ross, 2010). Their divergence was estimated to have occurred less than 200,000 years ago, coinciding with the rise of Mount Etna to elevations above 2,000 m due to volcanic activity (Chapman, Hiscock, & Filatov, 2013; Osborne, Batstone, Hiscock, & Filatov, 2013; Osborne, Chapman, Nevado, & Filatov, 2013). Both species are short-lived, obligately outcrossing perennials pollinated by generalist insects and bear wind-dispersed fruits; they can be distinguished by several morphological and physiological characters, such as degree of leaf dissection, size of capitula and florets, and flowering time (Brennan *et al.*, 2009; James & Abbott, 2005). The two species produce fertile hybrids (Chapman *et al.*, 2005), but significant hybrid breakdown in F_2 hybrids (Brennan, Hiscock, & Abbott, 2014; Chapman, Hiscock, & Filatov, 2016; Hegarty et al., 2009), and evidence of genetic incompatibilities (Brennan *et al*., 2014; Brennan, Hiscock, & Abbott, 2016, 2019; Chapman *et al.*, 2016) suggest the two are distinct species that have evolved substantial reproductive isolation. Details of research studies for this system will be discussed in the subsequent thesis chapters.

While hybrids of *S. aethnensis* and *S. chrysanthemifolius* often exhibit hybrid breakdown (present in Brennan *et al.,* 2014; Chapman *et al.,* 2016; Hegarty *et al.*, 2009; absent in Ross, 2010), a stabilised hybrid species of the two, *Senecio squalidus*, is widespread in the UK. Phenotypic studies have shown that *S. squalidus* are either similar to one of the two parental species or transgressive between the two; while genetic studies using randomly amplified polymorphic DNA (RAPD) and inter-simple sequence repeats (ISSR) showed *S. squalidus* to be genetically intermediate between the two parental species (Abbott, Hegarty, Hiscock, & Brennan, 2010; James & Abbott, 2005). *S. squalidus*, also called Oxford Ragwort, is a short-lived, obligately outcrossing perennial species which originated from materials taken from Mount Etna, Sicily to the UK at the beginning of the eighteenth century (Harris, 2002). Harris (2002) believed that plant materials were initially brought to and cultivated in the Duchess of Beaufort's Garden in Badminton, then the Oxford Botanic Garden, sometime before 1719. Plants have been cultivated in the Oxford Botanic Garden ever since. Linnaeus described *S. squalidus* in 1753 using plants grown from seeds sent from Oxford (Harris, 2002; Walker, 1833). A few decades after the introduction of plants to the Oxford Botanic Garden, *S. squalidus* was reported to be growing on walls in Oxford (Sibthorp, 1794); its spread was aided by the later establishment of the railway system in the British Isles (Druce, 1927). Herbarium records reveal that the species started spreading northwards in the late nineteenth century, reached northern England between early to mid-twentieth century (Abbott, Brennan, James, & Forbe, 2009), and was found in the Central Belt of Scotland in the 1950s and 1960s (Kent, 1966).

2.6 Aims of research

This thesis aims to fill knowledge gaps regarding speciation and adaptation in the Etnean *Senecio* system and its hybrid species in the UK.

It is obvious that there is gene flow between *S. aethnensis* and *S. chrysanthemifolius* at present and is a case of speciation with gene flow. There is empirical evidence suggesting that the two species exchanged genes since the beginning of their divergence and that a secondary contact scenario was unlikely (Filatov*,* Osborne, & Papadopulos, 2016). However, no studies have considered secondary contact scenario and some crucial demographic features for speciation with gene flow (such as

heterogeneous gene flow) before. Besides, little is known about the amount of gene flow in the past and at present. Hence, *Chapter 3* was set to reconstruct the most likely demographic history of the system by considering many possible cases of speciation with gene flow, and also important demographic features that shaped the system.

Next, a bottom-up approach was used to look for signatures of divergent selection and positive selection in the genome for the first time in this system. To look for signatures of divergent selection and positive selection, cline analyses of highly differentiated SNPs and genome scans were utilised respectively in *Chapter 3* and *4*. Then, the strength of selection against hybrids, strength and age of putative selective sweeps were estimated; putative adaptation genes were also identified. These are to examine how widespread selection is in the genome. The results would provide insights into whether the strong selection on few genes or multifarious selection scenario was true for this system. These would also provide candidate genes which can be used for functional analyses in the future to investigate the effects of these adaptation genes on fitness in the contrasting environments on Mount Etna.

Lastly, a greenhouse crossing experiment was carried out to generate multiple F_2 hybrid families with different genetic backgrounds of *S. aethnensis* and *S. chrysanthemifolius* in *Chapter 5*. Phenotypic differences, including those resembling hybrid breakdown, among families were investigated. Phenotypic characters of the natural hybrid, *S. squalidus*, collected in various locations of the UK were used to compare differences between synthetic and natural hybrids of the same progenitor species. Comparing synthetic and natural hybrids would uncover more details of this hybrid speciation (such as the incompatibilities hybrids needed to overcome and whether the natural hybrid gained novel mutations). Although four other studies (Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009; Ross, 2010) have looked at hybrid breakdown in synthetic F_2 hybrids of the two Etnean species before, each of them only involved one parental cross and each study had different experimental design hence results are not directly comparable. Ross (2010) also did not observe any breakdown in synthetic F_2 hybrids. One of the aims of this thesis is to create multiple families of hybrids with parents of different genetic background under the same conditions, to make sure any discrepancies in the results among families would be due to genetic differences but not other artefacts.

Generating more F_2 families would also allow hybrid breakdown to be quantified in an organised way and to identify new breakdown phenotypes that might not be recorded in the previous studies.

Overall, this thesis aims to investigate the detailed demographic processes at different elevations, genome-wide signals of divergent selection and positive selection, hybrid fitness and hybrid speciation in the Etnean *Senecio* system. Nosil *et al.* (2009b) pointed out that although many studies have been done on the ecological factors (such as divergent selection) and non-ecological factors (such as genetic incompatibility) of speciation, few studies have explored the interactions between the two during speciation. Hopefully, this thesis will provide substantial empirical data on both and we can start to link them together for a more comprehensive picture of what happened during the speciation of the Etnean *Senecio.*

3. Strong divergent selection at multiple loci in two closely related species of ragworts adapted to high and low elevations on Mount Etna

Published as:

Wong, E. L. Y., Nevado, B., Osborne, O. G., Papadopulos, A. S. T., Bridle, J. R., Hiscock, S. J. H., & Filatov, D. A. (2020). Strong divergent selection at multiple loci in two closely related species of ragworts adapted to high and low elevations on Mount Etna. *Molecular Ecology*, **29**(2), 394-412. https://doi.org/10.1111/mec.15319

3.1 Preamble

Findings from this chapter have been published in the journal *Molecular Ecology*. In the nextRAD dataset, sample collection and DNA extractions were carried out by Owen Osborne and Alex Papadopulos. Leaf morphology data was provided by Dmitry Filatov. I analysed all the raw data and prepared the manuscript. Wordings in this thesis chapter are mostly identical to those in the published article. However, part of the introduction, tables, figures, supplementary tables and figures have been re-formatted and/ or re-numbered to fit this chapter.

3.2 Abstract

Recently diverged species present particularly informative systems for studying speciation and maintenance of genetic divergence in the face of gene flow. I investigated speciation in two closely related *Senecio* species, *S. aethnensis* and *S. chrysanthemifolius*, which grow at high and low elevations, respectively, on Mount Etna, Sicily and form a hybrid zone at intermediate elevations. I used a newly generated genome-wide single nucleotide polymorphism (SNP) dataset from 192 individuals collected over 18 localities along an elevational gradient to reconstruct the likely history of speciation, identify highly differentiated SNPs, and estimate the strength of divergent selection. I found that speciation in this system involved heterogeneous and bidirectional gene flow along the genome, and species experienced marked population size changes in the past. Furthermore, I identified highly-differentiated SNPs between the species, some of which are located in genes potentially involved in ecological differences between species (such as photosynthesis and UV response). I analysed the shape of these SNPs' allele frequency clines along the elevational gradient. These clines show significantly variable coincidence and concordance, indicative of the presence of multifarious selective forces. Selection against hybrids is estimated to be very strong (0.16 – 0.78) and one of the highest reported in literature. The combination of strong cumulative selection across the genome and previously identified intrinsic incompatibilities probably work together to maintain the genetic and phenotypic differentiation between these species – pointing to the importance of considering both intrinsic and extrinsic factors when studying divergence and speciation.

3.3 Introduction

Defining what constitutes a species and understanding how new species form are central questions in evolutionary biology. The idea that speciation is driven by adaptation to distinct environmental conditions has a long history going back to Darwin's concept of new species formation by means of natural selection. Although natural selection is the main driver of evolutionary changes, its role in creating barriers to gene flow is not well understood. Adaptation to distinct environments (divergent selection) may lead to differentiation in loci responsible for adaptation or even fixation of different alleles in isolated populations. This can generate a mosaic with different genomic regions having different extents of gene flow, and consequently, genetic divergence (Feder, Egan, & Nosil, 2012; Mallet, 2005; Teeter *et al.*, 2008; Wu, 2001). This effect allows evolutionary biologists to study genetic and phenotypic differences between species in relation to environmental or ecological change, and in areas where populations meet, such as hybrid zones. Hybrid zones, and more generally systems in which reproductive isolation is incomplete, are powerful tools for this purpose. This is because the basis of reproductive isolation and recombinants can be more easily detected in these systems without being obscured by post-speciation divergence.

Genetic and phenotypic trait differences between species or populations can be tracked along geographical clines. Parameters of clines are informative about the strength of selection and dispersal rate (Barton & Gale, 1993; Barton & Hewitt, 1985; Szymura & Barton, 1986; 1991), while cline coincidence and concordance inform us about selective pressures in the hybrid zone. For instance, both coincidence and concordance could suggest they are influenced by their respective gradients (whether identical or different) in a similar way (Barton & Gale, 1993; Kruuk, Baird, Gale, & Barton, 1999; Young, 1996); whereas discordant clines could suggest the presence of several external selective pressures (Durrett, Buttel, & Harrison, 2000). Tracking the change in allele frequency or trait values along the clines allow us to evaluate the nature, magnitude and relative importance of selection in shaping the hybrid zone, thus enable a deeper understanding of microevolutionary bases of local adaptation and reproductive isolation. Several studies have estimated the strength of selection in different systems (Table 3.1), with estimated selection coefficient ranging from as low as 0.0017 to over 0.7. Surprisingly, although interspecific hybridisation is common in plants, there is a lack of

selection estimates for plant hybrid zones [e.g., Antonovics & Bradshaw, 1970 (this study estimated selection using a different method compared to the one in this study); Brennan *et al.*, 2009]. Our study aims to fill this gap, providing a detailed analysis of selection maintaining an altitudinal hybrid zone in *Senecio* on Mount Etna, Sicily.

It is thought that the chance of local adaptation leading to speciation can be increased by either strong selection on a single trait or selection on a larger number of traits (stronger selection and multifarious selection hypotheses; Nosil *et al.*, 2009b). Although speciation with one or few traits have been well characterised, an increasing number of cases are reporting the presence of multifarious selection in different systems. For example, selection by both breeding pool acidity and predators in frogs (*Rana arvalis*; Egea-Serrano, Hangartner, Laurila, & Räsänen, 2014), selection on female oviposition preference and diapause initiation in butterflies (*Lycaeides* sp.; Gompert *et al.*, 2013), selection on various diapause life-history traits in flies (*Rhagoletis* sp.; Daroski & Feder, 2007), and selection on floral colour and other ecophysiological traits in monkeyflowers (*Mimulus aurantiacus*; Sobel, Stankowski, & Streisfeld, 2019; Stankowski, Sobel, & Stresifeld, 2017). In cases where there is multifarious selection, although selection on each trait might not be strong enough to cause speciation on its own, it is the cumulative effect of many selective agents that leads to strong reproductive isolation and divergence overall; genetic hitchhiking leading to correlated response of many genes could also be effective in driving divergence in the multifarious selection scenario (Nosil *et al.*, 2009a, 2009b).

In this study, I created clines for several genetic and phenotypic markers, and estimated the strength of selection against hybrids in an altitudinal hybrid zone between *S. aethnensis* and *S. chrysanthemifolius* on the southern side of Mount Etna, Sicily. Although there might be other clines on other sides of the mountain, it was observed that samples collected in this study is the most continuous along the elevational gradient. Results from this cline would be representative of the species, as the selective forces are hypothesized to be associated strongly with elevation, but not other factors such as directions the mountain face. I also reconstructed the history of their speciation and describe the interplay among demography, gene flow and selection in this system. Previous demographic studies in this system (Chapman *et al.*, 2013; Filatov *et al.*, 2016; Muir *et al.*, 2013; Osborne *et al.*, 2013) have shown the presence of low ongoing gene flow between the species. It was

also demonstrated that an allopatric phase during divergence of the two *Senecio* species was unlikely (Filatov *et al.*, 2016). Other studies have identified differentially expressed genes (Chapman *et al.*, 2013); investigated geographic clines in the Etnean *Senecio* system using phenotypic data, allozymes and simple sequence repeats, and estimated moderate selection against hybrids (0.02–0.41; Brennan *et al.*, 2009); and mapped hybrid breakdown to several quantitative trait loci (Chapman *et al.*, 2016). However, these studies have not directly tested for other demographic parameters (such as heterogeneous gene flow and population size change), patterns of gene flow within the hybrid zone, and the extent of divergent selection. In this study, I addressed these issues specifically using finer scale sampling throughout the geographical cline on the southern side of Mount Etna, combined with analysis of leaf morphology and genetic diversity based on reduced-representation genomic data ("nextRAD"; Russello, Waterhouse, Etter, & Johnson, 2015). I hypothesize that: (a) as reproductive isolation is incomplete, the two species would show demographic features of recent speciation with gene flow, such as heterogenous gene flow; (b) the system is under multifarious selection thus there would be many genetic markers under various strengths of divergent selection; and (c) there is strong cumulative selection against hybrids to maintain species divergence despite ongoing gene flow. To test these hypotheses, I: (a) reconstructed the likely speciation scenario in this system, specifically including hybrid groups and demographic models comprising parameters previously unexplored; (b) identified potential genomic targets of divergent selection and carried out cline analysis on them; and (c) estimated the strength of selection in the hybrid zone. These analyses allowed us to infer the patterns of gene flow in different parts of the hybrid zone and the extent of selection that is responsible for the maintenance of species identity and build-up of species divergence despite ongoing interspecific gene flow.

Table 3.1. Recent studies estimating selection in hybrid zones using the approach of Barton and Gale (1993), ranked from highest to lowest estimates of selection strength.

Common name	Species	No. of loci	Selection estimates (CI)	Reference
Selection estimated using $s = 8\sigma^2/w^2$ (where σ = dispersal rate and w = cline width):				
Salamander	*Ensatina eschscholtzii*	9	0.46 - 0.75 (NA)	Alexandrino *et al.*, 2005
Rainforest skink	*Carlia rubrigularis* (N & S lineages)	9	0.50 – 0.70 (NA) 0.22 – 0.49 (NA)	Phillips *et al.*, 2004
Rainforest lizard	*Lampropholis coggeri* (C & S lineages)	11	0.403 (0.106 – 0.653)	Singhal & Moritz, 2012
Swainson's thrush	*Catharus ustulatus ustulatus* & *C. u. swainsoni*	3	0.19 – 0.40 (NA)	Delmore & Irwin, 2014
European rabbit	*Oryctolagus cuniculus cuniculus* & *O. c. algirus*	28	0.20 (0.05 – 0.64)	Carneiro *et al.*, 2013
Morabine grasshopper	*Vandiemenella viatical* (2 chromosomal races)	10	0.197 (0.058 – 0.405)	Kawakami, Butlin, Adams, Paull, & Cooper, 2009
Yellow-rumped warbler	*Dendroica coronata coronata* & *D. c. auduboni*	2	0.18 (0.08 – 0.28)	Brelsford & Irwin, 2009
Ragwort	*Senecio aethnensis* & *S. chrysanthemifolius*	13	0.02 – 0.11 (NA)	Brennan *et al.*, 2009
Common shrew	*Sorex araneus* (2 chromosomal races)	9 3	0.001 – 0.110 (NA) 0.0003 – 0.003 (NA)	Polyakov, White, Jones, Borodin, & Searle, 2011
House mouse	*Mus musculus musculus* & *M. m. domestica*	7	0.028 – 0.049 (NA)	Macholn *et al.*, 2007
Red-backed fairy-wren	*Malurus melanocephalus melanocephalus* & *M. m. cruentatus*	102	0.007 (0.002 – 0.03)	Baldassarre, White, Karubian, & Webster, 2014
Selection estimated using $s = 4\sigma^2/w^2$:				
Crested Newt	*Triturus anatolicus* & *T. ivanbureschi*	49	0.11 (0.004 – 0.019)	Wielstra *et al.*, 2017
Italian/ House sparrow	*Passer italiae* & *P. domesticus*	4	0.062 (0.038 – 0.109)	Bailey, Tesaker, Trier, & Sætre, 2015
Common/ Spined toad	*Bufo budo* & *B. spinosus*	32	0.0017 (0.0001 – 0.004)	van Riemsdijk *et al.*, 2019
Selection estimated using $s = 3\sigma^2/w^2$:				
Stickleback	*Gasterostens aculeatus* (stream & anadromous ecotype)	7	0.097 (NA)	Vines *et al.*, 2016
Marine gastropod	*Littorina saxatilis* (crab & wave ecotype)	57 69	0.06 (0.005 – 0.32) 0.008 (NA)	Hollander, Galindo, & Butlin, 2015

3.4 Materials and Methods

3.4.1 Sampling

To infer the extent of genome-wide divergence between *S. aethnensis* and *S. chrysanthemifolius*, and the potential selective pressures acting on this system, I analysed data comprising reduced-representation nextRAD sequences (Russello *et al.*, 2015) and leaf dissection measurements. The same individuals were used in generating both data sets. A total of 192 individuals from 18 localities were sampled over an elevational gradient from 585 to 2,645 m on a transect on the southern side of Mount Etna. These include six typical *S. aethnensis* (above 2,000 m) localities, three typical *S. chrysanthemifolius* (below 1,000 m) localities and nine hybrid localities (Figure 3.1d); nine to 14 individuals were sampled from each locality (Table 3.2). Leaf material was collected and dried in silica gel for DNA extractions; while fresh leaves were collected from each individual on site and photographed on top of gridded paper for measurements.

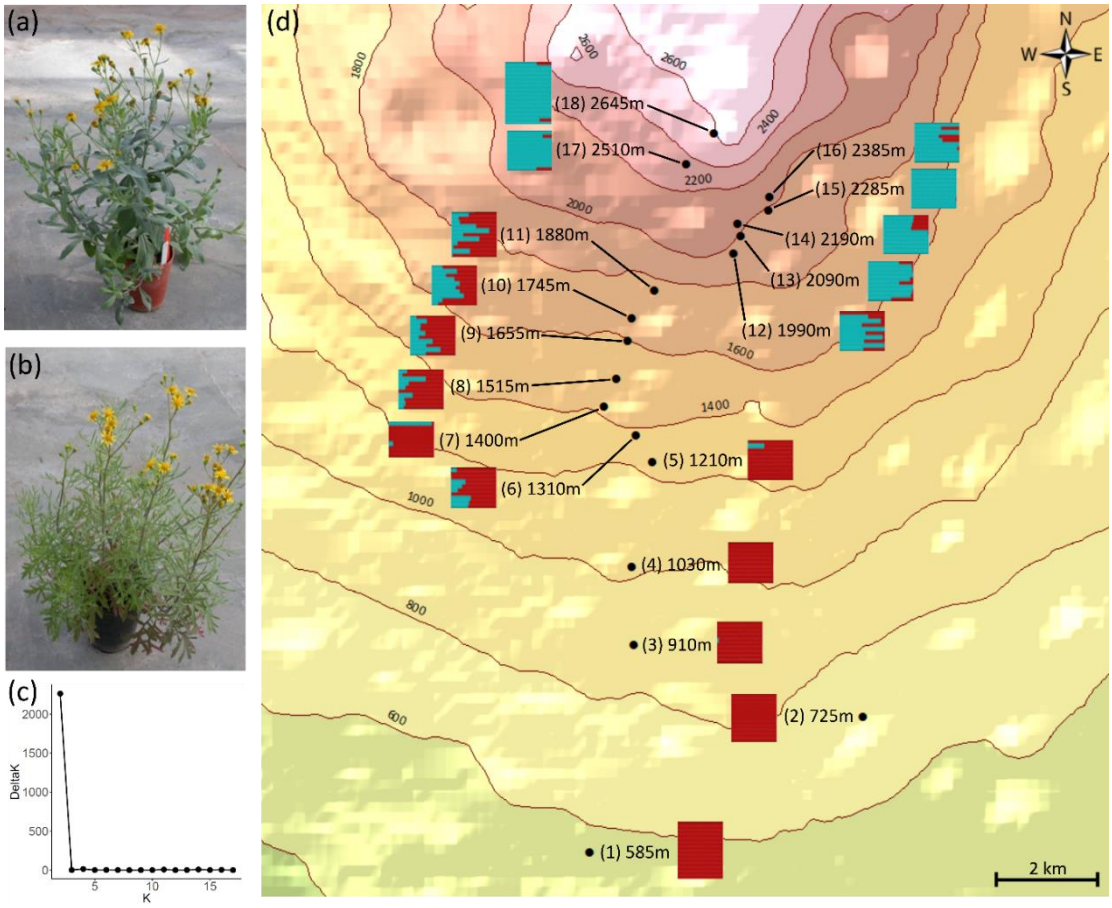


Figure 3.1. Distribution of sampled populations on Mount Etna and their genetic structure. (a) Photo of the high-elevation species *Senecio aethnensis* (obtained from senecioDB, undated). (b) Photo of the

low-elevation species *Senecio chrysanthemifolius* (obtained from senecioDB, undated). (c) Δ*K* plot showing optimal *K* = 2. (d) Contour map showing the southern side of Mount Etna using Google Earth Pro (Google) and ArcGIS (ESRI). Each black dot represents a sampled site. Numbers in brackets represent population numbers and numbers next to them are elevations where they were sampled. Coloured rectangular plots next to each site represent its corresponding section in the *STRUCTURE* plot (using 1,769 nextRAD markers with < 50% missing data and *K* = 2).

3.4.2 nextRAD sequencing

In the nextRAD data set, the two highest and lowest elevation localities were previously used in a demographic study (Filatov *et al.*, 2016), while the remaining 14 localities are new additions to the data set. Genomic DNA was extracted from dried leaves using a modified CTAB protocol (Doyle & Doyle, 1987) and purified with QIAGEN DNeasy mini-spin columns. DNA was then sent to SNPsaurus (Oregon, USA), who prepared nextRAD sequencing libraries and carried out 150 bp single-end Illumina sequencing. This approach is similar to RAD-seq, but uses degenerate primers instead of restriction enzymes.

3.4.3 Analysis of raw reads, genetic diversity and structure in nextRAD data set

Raw reads from nextRAD sequencing were processed using *STACKS* v1.34 (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011). The *process_radtags* function was used to remove any reads that contained uncalled bases and other low-quality reads. The *denovo_map.pl* script was then used to create a de novo catalogue of loci and genotype reads. A minimum stack depth (option -m) of 15, maximum number of mismatches between loci of the same individual (option -M) of two, and maximum number of mismatches between loci in the catalogue (option -n) of two were allowed in making the de novo catalogue. The *populations* function was used to call SNPs from the samples, calculate genetic diversity statistics (number of private alleles, percentage of polymorphic loci, observed heterozygosity, expected heterozygosity, nucleotide diversity, inbreeding coefficient, and pairwise F_{ST}), and create genotype input files for further analyses using other software packages. SNPs with 50% or less missing data were retained for subsequent analyses. Isolation by distance was assessed using pairwise genetic distance ($F_{ST}/[1-F_{ST}]$) and difference in elevation or geographic

distance between locality pairs, with Mantel test, all carried out in *IBD* v1.52 (Bohonak, 2002). Analysis of molecular variance (AMOVA), using only localities involved in demographic modelling, was carried out to test that polymorphism is mostly within species or hybrid groups instead of among individual localities (see demographic modelling section below for details on grouping). To assess the genetic clustering within the data set, I carried out a principal component analysis (PCA) with the R package *ade4* (Bougeard & Dray, 2018; Chessel, Dufour, & Thioulouse, 2004; Dray & Dufour, 2007; Dray, Dufour, & Chessel, 2007) and *factoextra* (Kassambara & Mundt, 2020). Missing data in PCA were replaced with mean allele frequencies. I also used *STRUCTURE* (Pritchard, Stephens, & Donnelly, 2000) to analyse the clustering of samples. In this analysis, the admixture model was used and the data was run for K = 2 to 18, with 10 iterations for each K. 200,000 generations of burnin was performed, with 100,000 generations retained. The value for K was then determined using the ad hoc statistic ΔK (Evanno, Regnaut, & Goudet, 2005) calculated with *Structure Harvester* v0.6.94 (Earl & vonHoldt, 2012).

3.4.4 Demographic modelling using nextRAD data

To explore the demographic history of the two *Senecio* species and their hybrids and visualise the two-dimensional site frequency spectra (2D-SFS), I investigated 14 demographic models (Figure 3.2) using Poisson random field-based demography inference framework implemented in the *dadi* package (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). For this analysis, localities were pooled to represent each of the typical species localities and hybrid localities at the top and bottom of the hybrid zone (inferred from *STRUCTURE* plot): localities 1 – 3 (585 – 910 m) were pooled to represent typical *S. chrysanthemifolius* (C); localities 7 – 9 (1,310 – 1,515 m) were pooled to represent low-elevation hybrids (LH); localities 12 – 14 (1,880 – 2,090 m) were pooled to represent high-elevation hybrids (HH); localities 16 – 18 (2,386 – 2,645 m) were pooled to represent typical *S. aethnensis* (A). This pooling strategy was necessary to increase the number of samples in each group. To avoid artificially inflating gene flow between groups through pooling localities, AMOVA was carried out to test that polymorphism was mostly within these groupings instead of among individual localities. Moreover, a test run using only one locality of each typical species (locality 1 and

18) were consistent with those from pooled localities. Since no outgroup data was available, folded site frequency spectra (SFS) were generated for each group using the python script *easySFS.py* (accessed from https://github.com/isaacovercast/easySFS). Following recommendations in the *dadi* manual, each group's sample size was projected down to 10 – 14 alleles (i.e., SNPs with less than these numbers of alleles scored will be removed, and those with more were subsampled to these numbers) to deal with missing data across individuals. SFS of the species pair was first analysed under 14 models of isolation with migration and secondary contact (Figure 3.2), which were published previously (Filatov *et al.*, 2016; Nevado, Contreras-Ortiz, Hughes, & Filatov, 2018; Tine *et al.*, 2014) or modified from them. These models vary in complexity (see detailed flow chart in Figure 3.2), with the simplest one (*split*) having no population size change and no migration between populations since splitting; to more complex models that allow migration rate in one direction only (e.g., *IM1*, *IM2*), same migration rate for both directions (e.g., *split_mig*, *IM2M_1*), different migration rates in each direction (e.g., *SC*, *IM*, *IMpre*), different migration rates at different times (e.g., *SplitExpMig*, *eSplitExpMig*) or at different parts of genome (e.g. *IM2M*) and population size fluctuations (e.g., *eSplitExpMig*, *IM*, *IMpre*, *IM2M*). Several models were used in this study to test conditions that were not explored before, such as unidirectional gene flow (*IM1*, *IM2*), heterogeneous gene flow (*IM2M*), and a secondary contact scenario with population size change (*eSC*, *SC_IM2M*). Since the models *IM2M* (heterogeneous gene flow since divergence) and *SC_IM2M* (period of no gene flow followed by heterogeneous gene flow after secondary contact) both have high likelihoods, a more complex model comprising features of both, *IM2M_AL_SC*, was included. This model assumes speciation with heterogeneous gene flow, followed by a period of no gene flow, then secondary contact with heterogeneous gene flow. For all models, 10 runs were performed initially for each data group using a wide parameter range (0–5 for time parameters, 0–10 for migration parameters, 0–100 for size parameters). A further set of 30 runs were then performed using narrower parameter ranges around the optimal values identified in the initial runs.

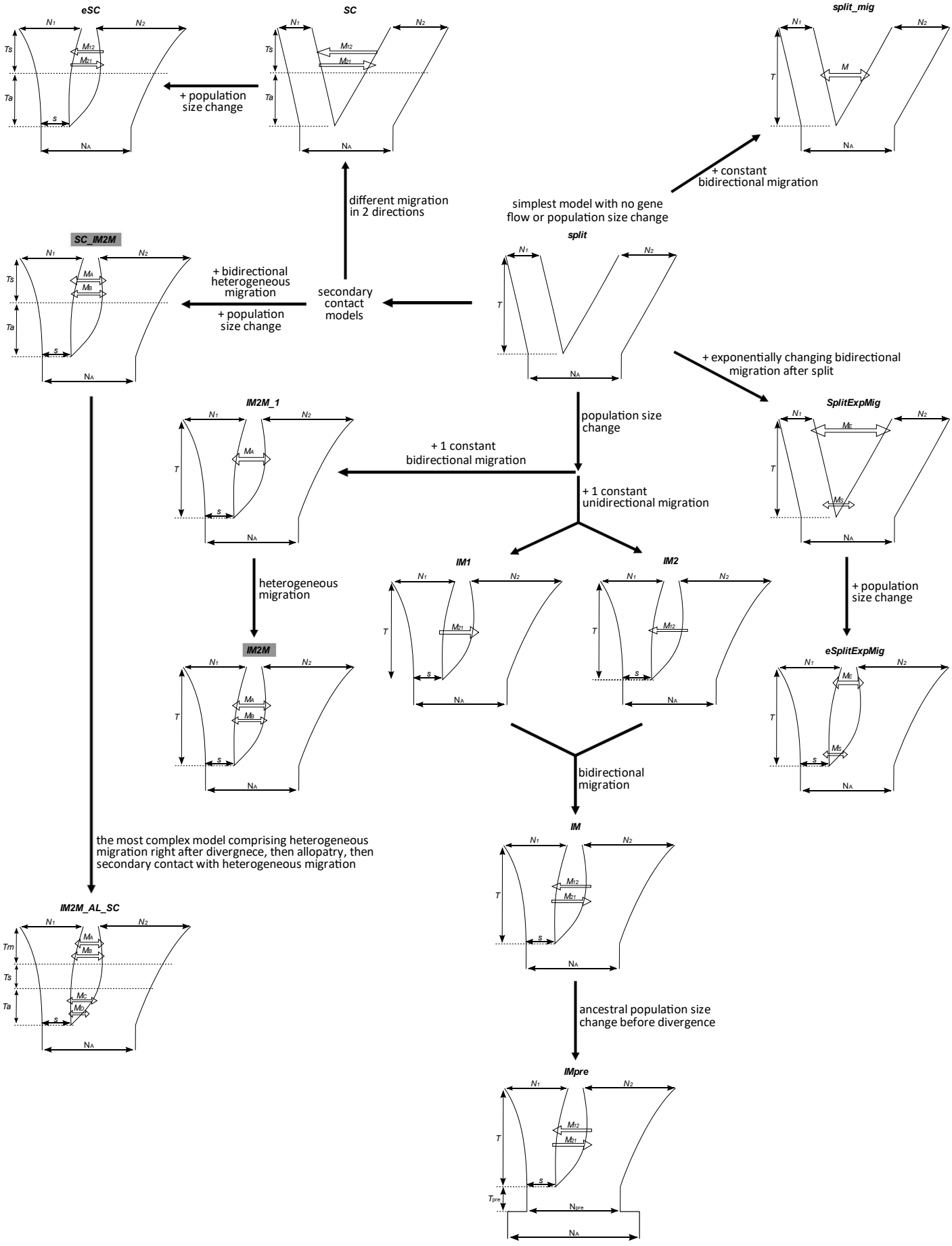

eSC
SC
split_mig
+ population size change
+ constant bidirectional migration
simplest model with no gene flow or population size change
different migration in 2 directions
SC_IM2M
split
+ bidirectional heterogeneous migration
+ population size change
secondary contact models
+ exponentially changing bidirectional migration after split
IM2M_1
population size change
+ 1 constant bidirectional migration
SplitExpMig
+ 1 constant unidirectional migration
IM1
IM2
heterogeneous migration
IM2M
+ population size change
eSplitExpMig
bidirectional migration
IM
the most complex model comprising heterogeneous migration right after divergnece, then allopatry, then secondary contact with heterogeneous migration
IM2M_AL_SC
ancestral population size change before divergence
IMpre

Figure 3.2. Flow chart showing extra parameters added to more complex models compared to simpler ones. Schematic representation of the demographic models is adapted from Nevado *et al.*, 2018. In each diagram, the top shows the two current population sizes (N_1 and N_2), and moving down (backward in time) the assumed demographic history since the split. Models that allow for exponential population size changes since the split have curved lines and include the parameter s, which is the relative size of population 1 just after the split (relative size of population 2 just after the split is 1 - s). Models that assume constant population size since the split are represented by straight lines, and if the sum of the two current population sizes (N_1+ N_2) is different from the ancestral size (N_A), then an instant population size change is assumed to occur during the split. For all models, N_A is the ancestral population size before the split, and is not a free parameter. All time parameters (T, T_a, T_s, T_{pre}) are given in units of 2*N_A generations. All population sizes (N_1, N_2 and N_{pre}) are expressed in units of N_A. All parameters related to time are given in units of 2*N_A generations. All migration parameters (M) are represented by horizontal arrows, and expressed in units of 2*N_A*m, where m is the proportion of the receiving population consisting of immigrants in each generation. Model *IM2M*, *SC_IM2M* and *IM2M_AL_SC* assumes two classes of sites which experience different migration rates, and include an extra parameter (P) not shown in the figures that denotes proportion of sites that experience a migration rate of M_A or M_C (proportion of sites that experience M_B or M_D is 1 - P). Names of the two highest likely models are highlighted in grey.

Since not all the models were nested, the best-fitting model was selected based on Akaike Information Criterion (AIC). The best-fitting model was chosen based on the lowest AIC score. Robustness of parameter estimates of the two best-fitting models was evaluated with 100 bootstrap runs, with the confidence intervals calculated as $M \pm 1.96X$ (where M is the likelihood parameter estimate and X is the standard deviation from Godambe Information Matrix (GIM) across replicates). Likelihood ratio tests were performed for nested models to assess which features of models were important in the demographic history of *Senecio* on Mount Etna (refer to Figure 3.2 for visualization of models). In particular, *split_mig* and *split* were used to test whether migration has been important; *IM* and *IM1* or *IM2* were used to test whether migration in only one direction would better fit the data; *IM* and *IMpre* were used to test whether there was population size change before the split. *IM2M_1* and *IM2M* were used to test whether having two classes of migration rates provided significantly better fit.

After ranking the demographic models for the pair of pure species, the best-fitting model (*IM2M*) and another simpler isolation with migration model (*IM*) were fitted to data from five other comparisons involving pure species and hybrids: high- and low-elevation hybrids (HH/LH), *S. chrysanthemifolius* and low-elevation hybrids (C/LH), *S. chrysanthemifolius* and high-elevation hybrids

(C/HH), *S. aethnensis* and low-elevation hybrids (A/LH) and *S. aethnensis* and high-elevation hybrids (A/HH). This is to determine finer scale demographic patterns between species and hybrid localities at different elevations and to compare gene flow in each direction over various spatial separation. These two models have the same type of migration parameters (heterogeneous gene flow or two unidirectional migration) as *SC_IM2M* (the other highly likely model) and *eSC*; they were run since they have fewer parameters than the latter two while yielding the same type of migration data.

3.4.5 Measurement of leaf phenotype

Leaf dissection is the most morphologically distinct character between *S. aethnensis* and *S. chrysanthemifolius*, which have undissected and highly dissected leaves respectively. Hybrids show gradual changes in leaf dissection along the elevation gradient, suggesting that this trait is under divergent selection. To quantify the differences in leaf dissection between the typical species and their hybrids, leaf area to perimeter ratio was used as a proxy for degree of dissection (same trait used in Brennan *et al.*, 2009; 2016). Leaf area to perimeter ratio was calculated from leaf photographs taken at the time of collection. Four to 10 fresh leaves from each individual were collected to control for within individual variability in leaf size. They were photographed on top of gridded paper for measurements. Leaf areas and perimeters were measured using the software *ImageJ* (Schneider, Rasband, & Eliceiri, 2012). Leaf area to perimeter ratio was then calculated for each leaf and averaged across leaves for each individual.

3.4.6 Identification of outlier loci in the nextRAD data set

Fixation index F_{ST} was calculated for each SNP across all samples using *Arlequin* (Excoffier, Laval, & Schneider, 2005), and P-values were adjusted for false discovery rate (Benjamini–Hochberg procedure). Negative F_{ST} was treated as zero. Another Bayesian-based outlier scan was carried out using *Bayescan* (Foll & Gaggiotti, 2008) to estimate the probability of a SNP being under selection. This analysis was carried out using default settings: 20 pilots runs of 5,000 iterations, followed by 50,000 burnin iterations and 5,000 iterations using a thinning interval of 10. Prior odds was set to 10,000. SNPs were treated as outliers if they were significant in both analyses – markers with adjusted P-value

smaller than 0.05 in *Arlequin* and $\log_{10}$PO larger than one (which indicates selection is 10 times more likely than neutral differentiation at the SNP) in *Bayescan*. Although F_{ST} is a relative measure (hence chance of resulting in false positives), it was used in detection of outliers as most widely available and used softwares are F_{ST}-based. The use of both *Arlequin* and *Bayescan*, and selecting only common outliers between the, aims at enhancing the confidence of the resulting outliers being true positives.

To test if the SNPs are part of or in proximity to any functional genes, a blast search of the loci containing the outlier SNPs was first carried out against genomic scaffolds of *S. squalidus* (unpublished; *S. squalidus* is the hybrid species between *S. aethnensis* and *S. chrysanthemifolius* that originated ex situ in the UK). Only non-duplicated matches with at least 95% base pair and 90% identity matched were retained. Sequences within 10 kb from the markers (total of 20,151 bp) were then blasted against the non-redundant protein sequences database of NCBI. Gene ontology (GO) terms of top blast hit genes were identified using *Uniprot* (https://www.uniprot.org/). 10 kb of sequences from the markers, but not more, from the markers were used in the blast search as a conservative and cautious threshold.

3.4.7 Cline fitting

Maximum likelihood fitting of clines was done to leaf dissection means or mean allele frequency along the elevational gradient for the leaf area to perimeter ratio, and each of the 76 outlier markers. Elevation is used instead of geographic distance because (a) genetic distance has a stronger correlation with difference in elevation than geographic distance (Figure 3.3); (b) environmental or ecological selective pressures probably change with elevation but not geographic distance; and (c) it allows direct comparison of cline centre and widths at particular elevations in different studies in the system, whereas use of geographic distance would cause confusion as different transects might start and end at different locations. Cline fitting was carried out using the R package *hzar* (Derryberry, Derryberry, Maley, & Brumfield, 2014). Average leaf area to perimeter ratio of each locality was calculated from the values obtained from *ImageJ*; input allele frequencies of the nextRAD markers were obtained using *STACKS*. All clines were fitted with three maximum-likelihood models: model I: p_{min}/p_{max} set to observed values without fitting exponential decay curves; model

II: p_{min}/p_{max} estimated without fitting any decay curves; and model III: p_{min}/p_{max} estimated with decay curves at both ends fitted (p_{min} and p_{max} denote the lowest and highest average allele frequency in a locality detected for a marker, respectively). The model with the highest likelihood based on AIC scores was chosen for each marker. Cline centre, cline width, minimum and maximum allele frequency observed in each cline, and their two log likelihood (2InL) support limits, were obtained using the same R package. From these measures, maximum change in allele frequency for each marker (Δp) and cline slope (Δp/width) were calculated.

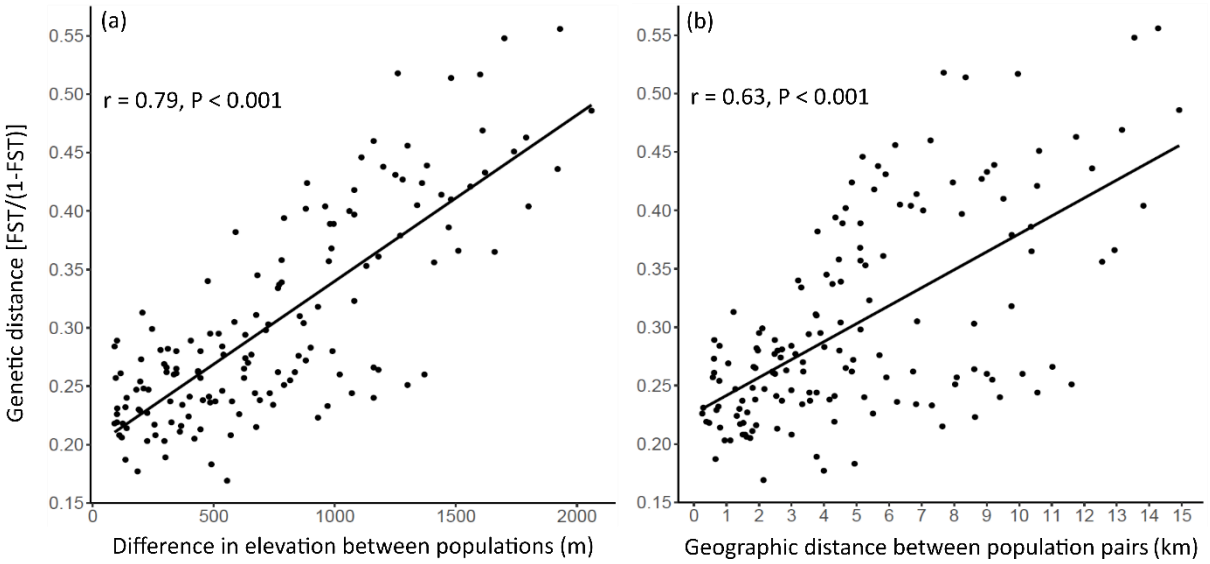


Figure 3.3. Mantel test for genetic distance, $F_{ST}/(1-F_{ST})$ and (a) difference in elevation between population pairs; (b) geographic distance between population pairs.

An "average" cline for molecular data was fitted using *STRUCTURE* ancestry scores for tests in cline coincidence (centre) and concordance (width); coincident or concordant clines could mean similar environmental gradient was underlying them. Centres or width of all outlier clines and leaf dissection cline were refitted and fixed to those of the *STRUCTURE* cline. The decrease in 2InL after fixing centre or width of each cline were tested using likelihood ratio tests, Bonferroni corrected for multiple tests. Clines were regarded as displaced or discordant with the *STRUCTURE* cline if the decrease in 2InL was significant. Likelihood profiles for selected markers covering the observed range of cline centres and widths were also created (following Phillips, Baird, & Moritz, 2004).

3.4.8 Strength of selection

A modified R script (original script provided by van Riemsdijk, Butlin, Wielstra, & Arntzen, 2019) was used to estimate the linkage disequilibrium, dispersal rate, and strength of selection in the hybrid zone. The average dispersal distance per generation (σ) was obtained by $\sqrt{(rDw^2)}$, where r is recombination rate, D is linkage disequilibrium at hybrid zone centre, w is mean cline width of markers used; effective selection against hybrids (s) was obtained by $4\sigma^2/w^2$ (Barton & Gale, 1993; Barton & Hewitt, 1985; Szymura & Barton, 1986, 1991). Since recombination rate in the species is not known, a range of 0.1 to 0.5 was used. Forty markers with coincident and concordant clines were used in the estimation of selection (the biggest group of coincident and concordant clines in the data set). These markers include the 33 markers in the elevational range of 1,900 – 2,000 m and with width around 1,200 m (as shown in Figure 3.11) and those that are coincident and concordant with them (tested with LRT, using the average centre and width of the group of markers at 1,900 – 2,000 m). Mean and 95% confidence intervals (CI) for linkage disequilibrium and effective selection were calculated with 100 iterations of bootstrapping with replacement.

3.4.9 Effect of selection on wider genomic regions

To test whether the signal of selection in the outlier markers (with annotated adjacent genes) was detectable in a wide genomic region, non-outlier SNPs (contained in loci that did not show duplication in blast searches, with 95% base pair and 90% identity matched with genomic scaffolds) closest to the outlier SNPs on the same genomic scaffold were identified and their F_{ST} values were compared to the outlier SNPs.

3.5 Results

To analyse genetic diversity and divergence among *Senecio* on Mount Etna, I conducted "nextRAD" sequencing for a sample of 192 individuals from 18 localities across an elevational gradient on Mount Etna (Figure 3.1d). In total, I generated 4.79×10^8 sequence reads that were mapped across 513,036 loci identified by *STACKS* (Catchen *et al.*, 2011). Most of these loci (413,124) were invariable, while 99,427 loci contained at least one SNP. Out of these loci, 1,769 SNPs with <50% missing data were used for analyses.

3.5.1 Genetic diversity and structure of localities inferred from nextRAD data set

Various genetic diversity indices showed substantial variation (Table 3.2, Figure 3.4). In general, *S. chrysanthemifolius* and low-elevation hybrid localities showed higher diversity than *S. aethnensis* and high-elevation hybrid sites. *S. chrysanthemifolius* and low-elevation localities showed generally a higher number of private alleles, a higher percentage of polymorphic loci, higher nucleotide diversity, a lower inbreeding coefficient (F_{IS}), higher observed heterozygosity and lower observed homozygosity (Table 3.2, Figure 3.4). Pairwise F_{ST} increased with difference in elevation and geographic distance between localities (Table 3.3).

After excluding SNPs with more than 50% missing data, 1,769 SNPs were retained for subsequent analyses. In the isolation by distance analysis, genetic distance had a stronger correlation with difference in elevation (Mantel test for genetic distance $F_{ST}/[1-F_{ST}]$: $r = 0.79$, $P < 0.001$ in Figure 3.3; pairwise F_{ST} matrix in Table 3.3) than geographic distance between locality pairs ($r = 0.63$, $P < 0.001$). Analysis of molecular variance (AMOVA) showed that most genetic variation is found among species or hybrid groupings (37.05%) and among individuals (58.98%) (Table 3.4). *STRUCTURE* and subsequent analyses revealed the optimal number of clusters (*K*) is two (Figure 3.1c; the same data set but with the 76 SNP outliers excluded showed the same results thus results from the full 1769-SNP data set was used in all subsequent analyses). Most individuals at the two ends of the elevational gradient clustered into two distinct groups, each presumably representing

one typical ("pure") species; admixed individuals were found at intermediate elevations, with a varying proportion of each cluster. There is a small amount of admixture detected in very high-elevation *S. aethnensis* samples, while very low-elevation *S. chrysanthemifolius* does not show any admixture. Principal components analysis (PCA; Figure 3.5) does not show conspicuous clustering of populations, but a gradual transition between high and low elevation samples. PC 1 shows high-elevation samples being more evenly distributed across the principal component while low-elevation ones are more clustered with several outliers in high-elevation samples' space. PC 2 shows rough clustering of each of the pure species and hybrids.

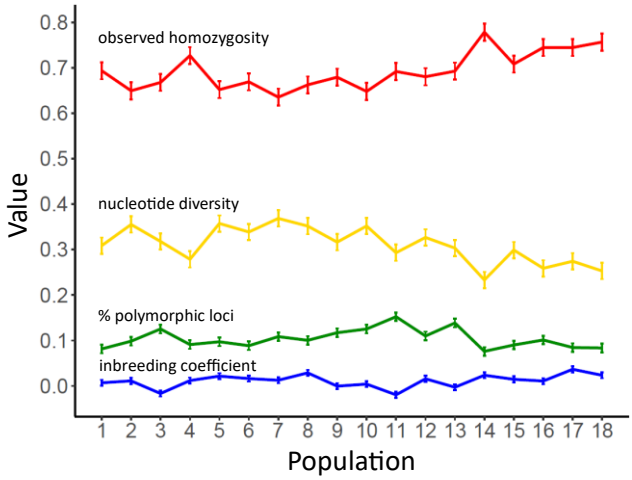


Figure 3.4. Change in five genetic diversity statistics across populations on Mount Etna, based on nextRAD data. Red line represents observed homozygosity; yellow line represents nucleotide diversity (π); green line represents percentage of polymorphic loci; blue line represents inbreeding coefficient (F_{IS}). Error bars show 95% confidence intervals.

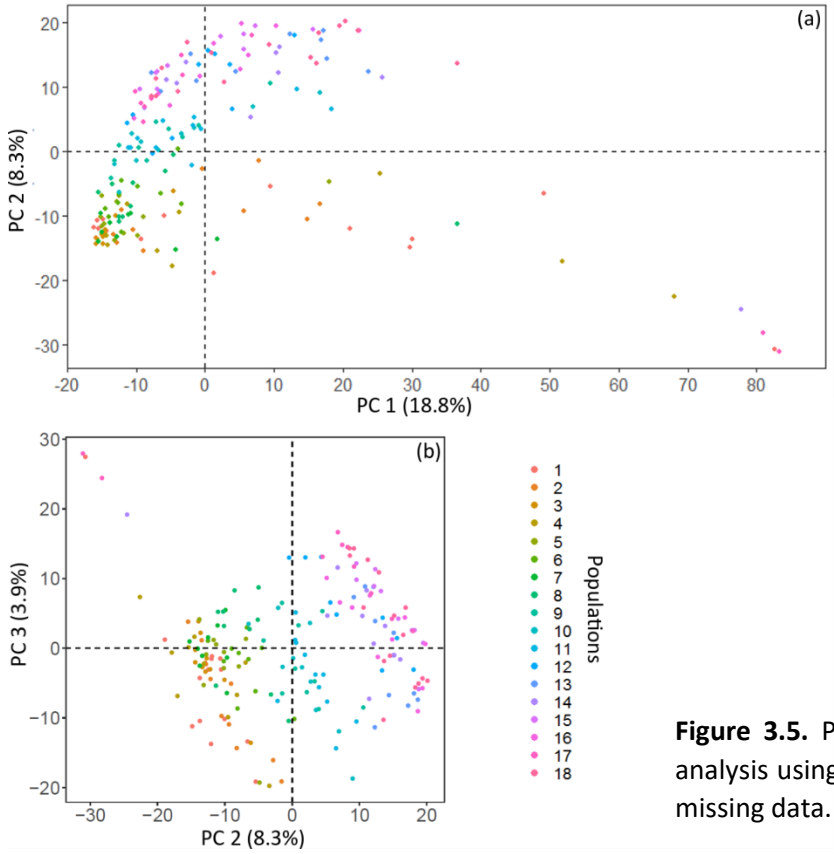


Figure 3.5. Plots from principle component analysis using 1,769 SNPs with less than 50% missing data. (a) PC 1 and 2. (b) PC 2 and 3.

Table 3.2. Genetic diversity statistics for polymorphic loci in each population.

Pop.	N	Altitude (m)	No. of private alleles	% Polymorphic Loci	H_o	Variance	SE	H_e	Variance	SE	π	Variance	SE	F_{IS}	Variance	SE
1	14	585	2025	0.0813	0.3062	0.1921	0.0029	0.1756	0.0521	0.0015	0.3086	0.1808	0.0028	0.0068	29.9966	0.0167
2	12	725	2383	0.0984	0.3504	0.2082	0.0027	0.1973	0.0550	0.0014	0.3552	0.1983	0.0027	0.0110	27.9420	0.0146
3	10	910	1825	0.1252	0.3321	0.1983	0.0029	0.1917	0.0534	0.0015	0.3179	0.1680	0.0027	-0.0168	27.9847	0.0162
4	10	1030	1471	0.0909	0.2730	0.1802	0.0030	0.1598	0.0499	0.0016	0.2787	0.1697	0.0029	0.0117	31.9099	0.0135
5	10	1210	2492	0.0971	0.3479	0.2095	0.0027	0.2002	0.0554	0.0014	0.3575	0.1972	0.0026	0.0215	27.8520	0.0130
6	10	1310	1950	0.0884	0.3308	0.2070	0.0028	0.1867	0.0550	0.0015	0.3387	0.1983	0.0028	0.0160	29.5107	0.0138
7	9	1400	2380	0.1084	0.3645	0.2155	0.0028	0.2085	0.0564	0.0014	0.3686	0.1971	0.0027	0.0124	27.0257	0.0122
8	10	1515	2514	0.1003	0.3374	0.2060	0.0027	0.1973	0.0551	0.0014	0.3518	0.1958	0.0026	0.0286	28.1924	0.0122
9	10	1655	1481	0.117	0.3206	0.1968	0.0031	0.1849	0.0533	0.0016	0.3164	0.1759	0.0030	-0.0006	29.1739	0.0153
10	10	1745	1570	0.1252	0.3519	0.2092	0.0031	0.1994	0.0555	0.0016	0.3519	0.1926	0.0029	0.0039	27.9976	0.0136
11	11	1880	1366	0.1522	0.3080	0.1759	0.0033	0.1894	0.0499	0.0017	0.2931	0.1427	0.0030	-0.0197	27.1403	0.0188
12	10	1990	2115	0.1098	0.3195	0.1991	0.0029	0.1846	0.0537	0.0015	0.3267	0.1874	0.0028	0.0156	29.3999	0.0130
13	10	2090	710	0.1384	0.3072	0.1925	0.0037	0.1763	0.0522	0.0019	0.3033	0.1742	0.0035	-0.0030	30.0069	0.0183
14	10	2190	860	0.076	0.2213	0.1562	0.0033	0.1342	0.0448	0.0017	0.2331	0.1516	0.0032	0.0234	34.4865	0.0167
15	10	2285	1218	0.09	0.2917	0.1926	0.0032	0.1643	0.0517	0.0017	0.2990	0.1873	0.0032	0.0144	31.7083	0.0166
16	10	2385	948	0.101	0.2553	0.1701	0.0033	0.1549	0.0485	0.0018	0.2588	0.1537	0.0031	0.0107	32.1974	0.0174
17	10	2510	1613	0.0844	0.2553	0.1703	0.0029	0.1583	0.0489	0.0015	0.2743	0.1663	0.0029	0.0365	31.9609	0.0134
18	10	2645	1652	0.0834	0.2432	0.1623	0.0028	0.1529	0.0471	0.0015	0.2529	0.1488	0.0027	0.0233	31.7853	0.0224

Table 3.3. Pairwise F_{ST} between each population pair.

Population	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.193	0.179	0.204	0.204	0.233	0.203	0.241	0.196	0.210	0.200	0.262	0.268	0.319	0.354	0.288	0.357	0.327
2		0.150	0.208	0.191	0.234	0.177	0.201	0.182	0.206	0.194	0.275	0.206	0.278	0.296	0.268	0.317	0.303
3			0.171	0.159	0.194	0.155	0.184	0.190	0.208	0.189	0.284	0.209	0.299	0.305	0.291	0.341	0.311
4				0.198	0.220	0.190	0.228	0.210	0.229	0.216	0.288	0.286	0.315	0.341	0.298	0.339	0.302
5					0.224	0.187	0.210	0.175	0.197	0.196	0.253	0.214	0.280	0.295	0.265	0.313	0.293
6						0.221	0.238	0.219	0.208	0.173	0.257	0.263	0.287	0.263	0.244	0.305	0.288
7							0.207	0.178	0.209	0.194	0.277	0.192	0.283	0.298	0.269	0.308	0.301
8								0.176	0.198	0.178	0.254	0.191	0.238	0.252	0.233	0.280	0.261
9									0.179	0.169	0.206	0.208	0.221	0.227	0.196	0.237	0.219
10										0.158	0.230	0.207	0.219	0.217	0.213	0.250	0.220
11											0.172	0.198	0.220	0.224	0.192	0.215	0.208
12												0.180	0.215	0.212	0.183	0.228	0.217
13													0.185	0.187	0.169	0.170	0.145
14														0.204	0.202	0.192	0.192
15															0.187	0.185	0.174
16																0.179	0.172
17																	0.188

Table 3.4. Distribution of molecular variation at the individual, population and species/hybrid group levels (AMOVA). Only individuals used in demographic modelling were included.

	df	Sum of squares	Variance component	% of variation
Among species/hybrid groups	3	25475.50	239.74	37.06
Among populations	8	5270.24	25.66	3.97
Among individuals	119	45403.84	381.55	58.98
Total	130	76149.586	646.949	100

3.5.2 Demographic history of *Senecio* on Mount Etna

I first focused on the analysis of demographic history between the pure species, as done in the previous work (Filatov *et al.*, 2016). That previous work focused on overall species demography and did not consider the possibility that different parts of the genome may have different rates of interspecific gene flow due to selection preventing introgression of some genomic regions. To take the possibility of heterogeneous gene flow across the genome into account, I allowed an *IM* model to have two bidirectional migration parameters, M_A and M_B (model *IM2M*, Figure 3.6a). The *IM2M* model fits data significantly better than the models without heterogeneous migration (AIC score = 399.04; Table 3.5, Figure 3.6a). Another model, *SC_IM2M*, was almost as likely (AIC score = 399.14; Table 3.5, Figure 3.6b). Both models involve heterogeneous gene flow, with the higher migration rates being 4.38 and 4.96 times larger than the other lower ones for the models *IM2M* and *SC_IM2M* respectively. Likelihood ratio tests (LRT) were carried out for five groups of nested models. The LRT between *IMpre* and *IM* models was not significant (P-value = 0.57), indicating that allowing for an ancestral population size change prior to species split (in model *IMpre*) does not significantly improve the fit to data. LRT for other four model comparisons were significant. In particular, allowing for migration dramatically improves the fit to data compared to the nested model without migration (*split_mig* vs. *split*, P-value = 1.76×10^{-30}). Allowing migration in both directions significantly improves the fit to data compared to unidirectional migration (*IM* vs. *IM1* or *IM2*, P-value = 5.55×10^{-13} and 3.74×10^{-16} respectively). Having two separate classes of bidirectional migration significantly improves model fit compared to only one class of migration (*IM2M* vs. *IM2M_1*, P-value = 0.0066). These results indicate that the genomes are experiencing gene flow in both directions and that gene flow is heterogeneous (i.e., a fraction of the genome is exchanging alleles significantly less than the rest of the genome). Although the models assuming heterogeneous gene flow did not test for symmetry of gene flow in each class, results from models with separate parameters for different directions of migration (*IM*, *IMpre*, *SC*, *eSC*; see Table 3.5) suggest that gene flow in the same class is slightly asymmetric. Migration rate from *S. chrysanthemifolius* to *S. aethnensis* is 1.24 to 1.70 times higher than that from *S. aethnensis* to *S. chrysanthemifolius*.

Table 3.5. The likelihood of each demographic model investigated in *dadi* for each species/species or species/hybrid pair, results of likelihood ratio test (LTR) for certain nested models, Akaike information criterion (AIC) scores, ΔAIC scores (relative to best model), and the relative likelihood of each model compared with the best model, and their respective parameter estimate.

Pops	Model	No. of free param.	*dadi*					LRT		N_1	N_2	T	s	P
			log-likelihood	theta	AIC									
					AIC	ΔAIC	Rel. likelihood	2ΔLL	P-value					
C/A	eSplitExpMig	6	-197.31	392.91	406.62	7.54	0.0230			0.20	0.25	0.83	0.29	
	SplitExpMig	5	-198.70	540.87	407.39	8.32	0.0156			0.15	0.22	0.99		
	split_mig	4	-198.61	616.79	405.23	6.15	0.0462			0.13	0.20	1.10		
	split	3	-264.45	367.97	534.91	135.83	0.0000	131.68*	$1.76x10^{-30}$	0.28	0.28	0.23		
	IMpre	8	-196.96	216.71	409.92	10.85	0.0044			0.38	0.42	4.05	0.34	
	IM	6	-197.52	476.12	407.03	7.96	0.0187	1.11	0.57	0.15	0.16	0.66	0.33	
	IM1	5	-223.52	404.69	457.04	57.96	0.0000	52.00*	$5.55x10^{-13}$	0.16	0.20	0.35	0.65	
	IM2	5	-230.70	358.62	471.40	72.33	0.0000	66.37*	$3.74x10^{-16}$	0.19	0.35	0.40	0.21	
	IM2M	***7***	***-192.54***	***433.13***	***399.07***	***0***	***1***			***0.20****(0.18-0.22)*	***0.21****(0.17-0.25)*	***0.70****(0.45-0.95)*	***0.14****(0.12-0.17)*	***0.42****(0.33-0.52)*
	IM2M_1	5	-197.56	411.37	405.13	6.05	0.0485	10.05*	0.0066	0.18	0.25	0.98	0.35	
	SC	6	-212.66	182.06	437.32	38.25	0.0000			0.41	0.52			
	eSC	7	-196.45	384.47	406.89	7.82	0.0201			0.18	0.13		0.30	
	SC_IM2M	***8***	***-191.57***	***372.96***	***399.14***	***0.07***	***0.9666***			***0.22****(0.22-0.23)*	***0.10****(0.10-0.10)*		***0.23****(0.22-0.24)*	***0.26****(0.20-0.32)*
	IM2M_AL_SC	11	-194.14	290.02	410.27	11.20	0.0037			0.25	0.23		0.27	0.23
C/LH	IM	6	-246.61	293.92						0.41	0.82	4.55	0.44	
	IM2M	7	-253.23	297.13						0.27	0.88	3.52	0.31	0.13
C/HH	IM	6	-194.68	382.58						0.26	0.23	0.50	0.18	
	IM2M	7	-198.61	366.80						0.26	0.27	0.42	0.07	0.43
A/LH	IM	6	-234.40	506.69						0.25	0.24	2.19	0.58	
	IM2M	7	-222.65	404.34						0.31	0.24	0.66	0.26	0.03
A/HH	IM	6	-173.99	333.27						0.44	0.37	1.02	0.42	
	IM2M	7	-165.41	355.74						0.39	0.33	0.37	0.10	0.11
HH/LH	IM	6	-219.19	397.19						0.30	0.33	1.43	0.31	
	IM2M	7	-218.61	358.97						0.32	0.34	0.71	0.41	0.64

Table 3.5 (continued)

Pops	Model	M_A	M_B	M_C	M_D	M_{12}	M_{21}	M_S	M_E	M	T_a	T_s	T_m	T_{pre}	N_{pre}
C/A	eSplitExpMig							0.41	0.75						
	SplitExpMig							0.93	1.01						
	split_mig									1.14					
	split									0†					
	IMpre					0.31	0.53							0.22	4.31
	IM					0.76	1.24								
	IM1					0†	1.03								
	IM2					0.62	0†								
	IM2M	***0.31****(0.29-0.33)*	***1.36****(0.12-0.15)*												
	IM2M_1	0.74	0†												
	SC					0.42	0.63				3.04	0.83			
	eSC					0.62	0.77				0.31	0.64			
	SC_IM2M	***2.23****(0.22-0.23)*	***0.45****(0.42-0.48)*								***0.38****(0.37-0.39)*	***0.46****(0.44-0.48)*			
	IM2M_AL_SC	1.93	0.45	2.34	0.16						0.18	3.86	4.25		
C/LH	IM					1.73	3.74								
	IM2M	1.47	3.31												
C/HH	IM					0.77	2.33								
	IM2M	0.86	1.97												
A/LH	IM					1.28	2.01								
	IM2M	0.98	3.41												
A/HH	IM					1.73	4.09								
	IM2M	0.55	4.70												
HH/LH	IM					3.72	3.04								
	IM2M	3.61	2.26												

Remarks: under the 'Pops' column: C = *S. chrysanthemifolius*; LH = low-elevation hybrids; HH = high-elevation hybrids; A = *S. aethnensis*; the two best-fitting models are in bold and italic; in all cases, likelihood ratio tests are for the comparison with the model immediately above except for IM2, which is for comparison with IM; Under 'LRT', * indicates significant value; †indicates fixed value of parameters. Values in brackets for the models *IM2M* and *SC_IM2M* are confidence intervals obtained from bootstrapping. Refer to Figure 3.2 for definitions of parameters.

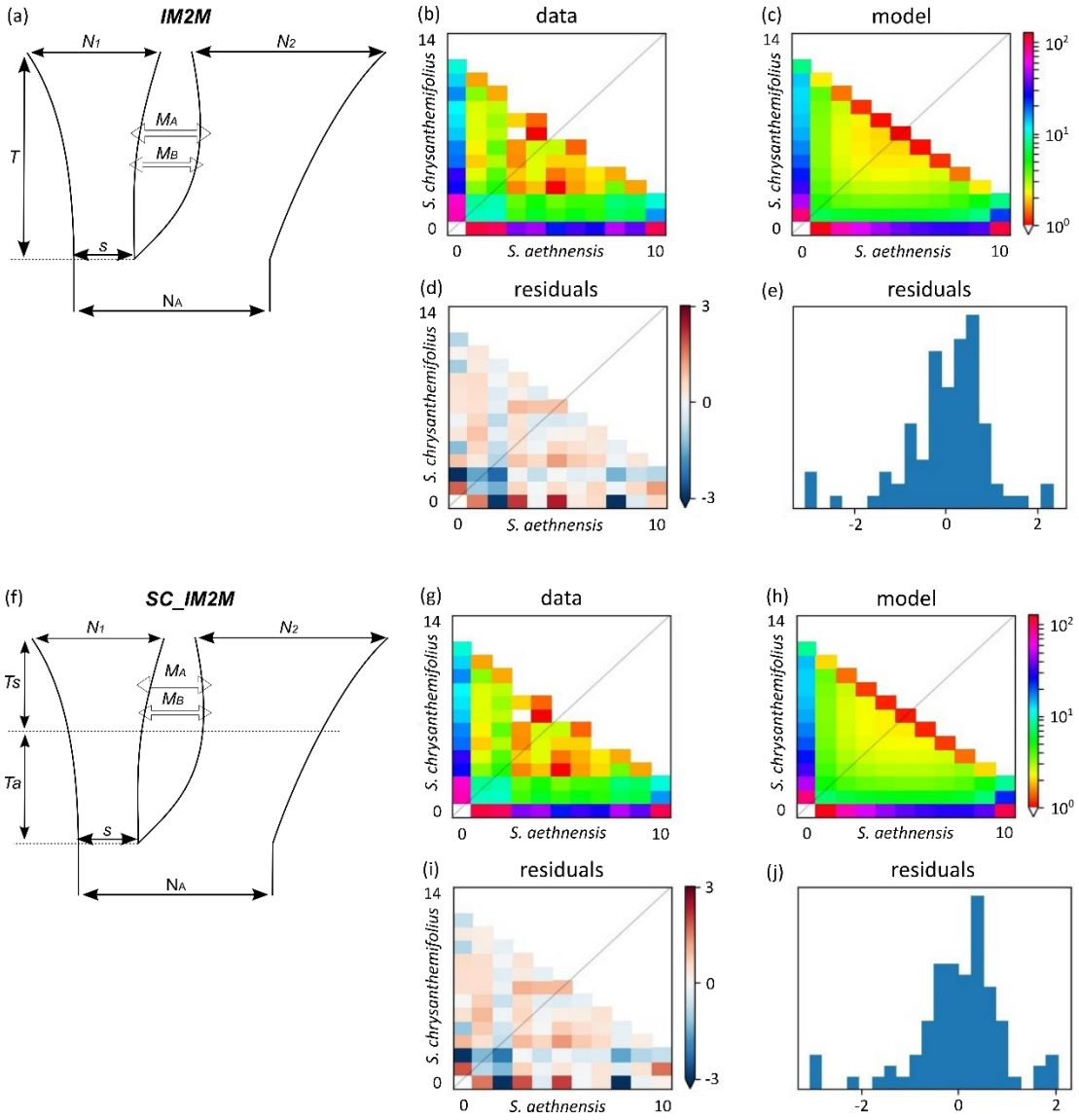


Figure 3.6. The fit of the two best demographic model (*IM2M* and *SC_IM2M*) to site frequency spectrum data for *S. aethnensis* and *S. chrysanthemifolius*. (a, f) Schematic representation of the two best demographic model. Each model has four corresponding figures each: (b–e) for *IM2M* and (g–j) for *SC_IM2M*. (b, g) Observed two-dimensional site-frequency-spectrum (2D-SFS). (c, h) 2D-SFS expected under the respective model. (d–e, i–j) Residuals between the observed and the expected site-frequency-spectra.

The best fitting model (*IM2M*) was also fitted to data from other pairwise comparisons of species and hybrid groups to investigate different demographic parameters at a finer scale. Another isolation with migration model, *IM*, was also run to assess the relative amount of gene flow in each direction in different comparisons (see parameters in Table 3.5). All groups show smaller current population sizes compared to their ancestral sizes. In all species/hybrid comparisons (C/LH, C/HH, A/LH, A/HH, where C = *S. chrysanthemifolius*, A = *S. aethnensis*, LH = low-elevation hybrids and HH = high-elevation hybrids) under the model *IM2M*, one class of migration is always at least twice (2.3 to 8.5 times) larger than the other one. Migration is also higher between localities that are closer to each other (C/LH and A/HH) than those further apart (C/HH and A/LH). Under the *IM* model, the migration rate from the pure species to hybrids (M_{21}) is 1.6 to 3 times larger than that in the opposite direction (M_{12}).

In the high- and low-elevation hybrids comparison (HH/LH), the two classes of migration rate are both high under the model *IM2M* (3.61 and 2.26) and *IM* (3.72 and 3.04), probably indicating most of their genomes share genes more freely (instead of having regions with restricted gene flow).

3.5.3 Cline in leaf dissection

The change in leaf area to perimeter ratio shows a clinal pattern (Figure 3.7). The cline showed a rapid change in leaf area to perimeter ratio around the elevation of 1,900 m, with the cline centre estimated at 1,910 m (2InL = 1,882 – 1,935 m), and the cline width of 127 m (2InL = 51 – 196 m).

3.5.4 Identification of outlier nextRAD markers and analysis of their clines

To identify highly differentiated markers, I used *Arlequin* and *Bayescan*. The F_{ST} values, calculated for each SNP using *Arlequin*, ranged from 0 to 0.8587 (mean F_{ST} = 0.0854). A total of 183 SNPs were identified as outliers based on F_{ST} values (Figure 3.8). *Bayescan* identified 89 SNPs with high F_{ST} and $\log_{10}PO > 1$ (Figure 3.9), indicating they are likely to be under divergent selection.

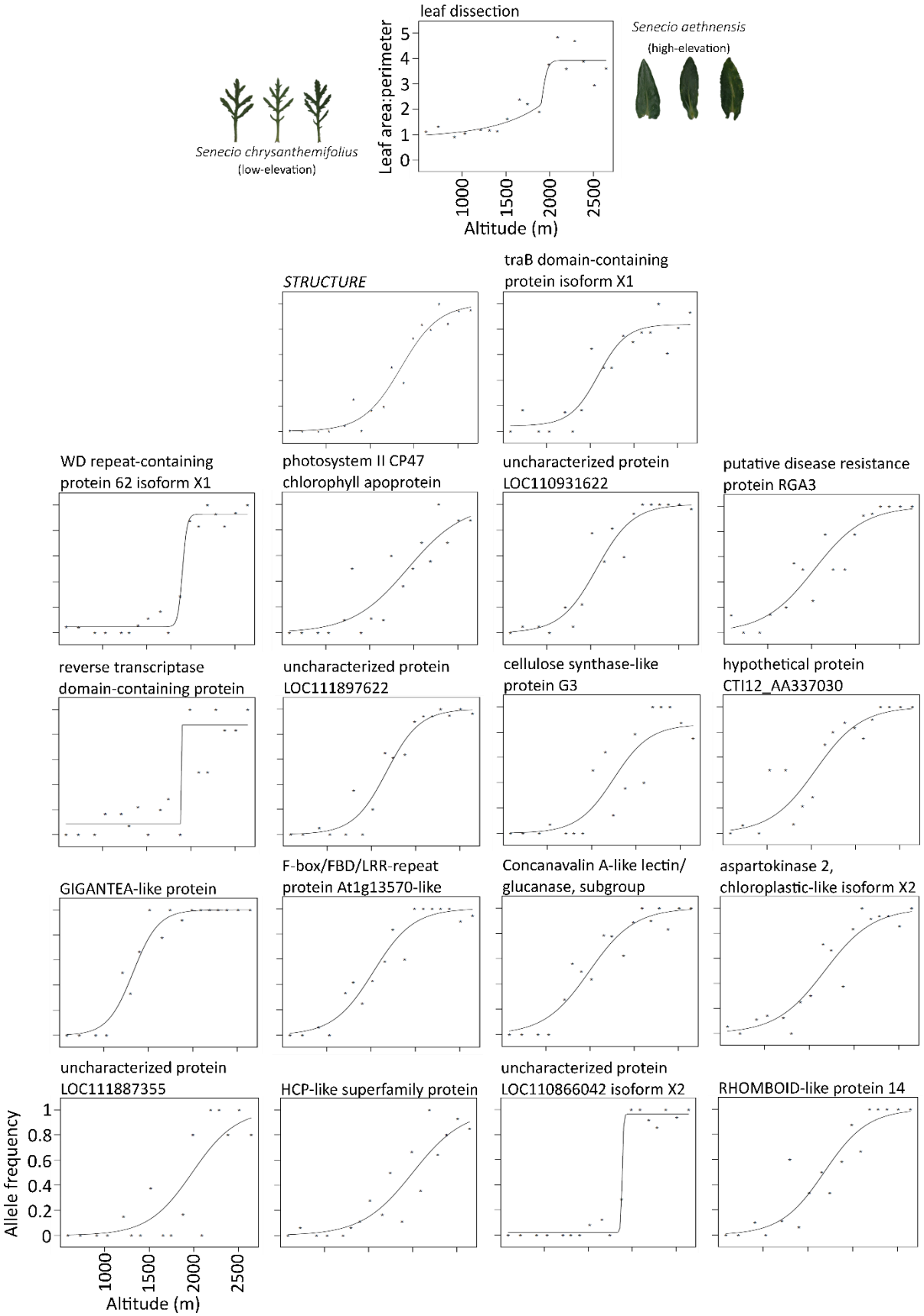


Figure 3.7. Clines for leaf dissection, *STRUCTURE*, and allele frequencies at 17 outlier SNPs located in genes with functional annotation available (see Table 3.7 for details). Each asterisk in the plots represents mean trait value, ancestry score, or allele frequency for each population. Examples of leaf shape of each species are shown next to the leaf dissection cline.

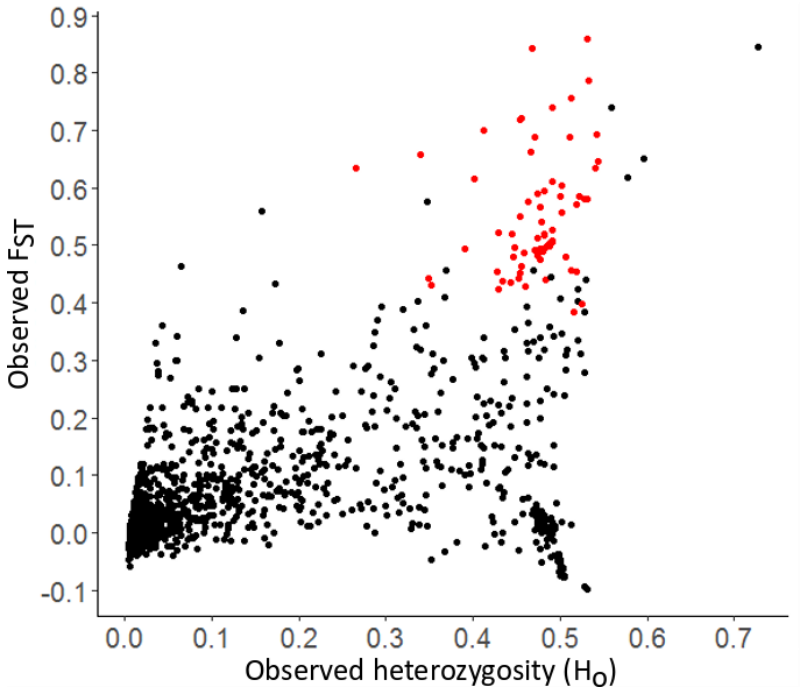


Figure 3.8. Species differentiation (F_{ST}) and observed heterozygosity (H_O) in 1,769 SNPs analysed by *Arlequin*. Red dots represent the 76 outlier SNPs identified in this study.

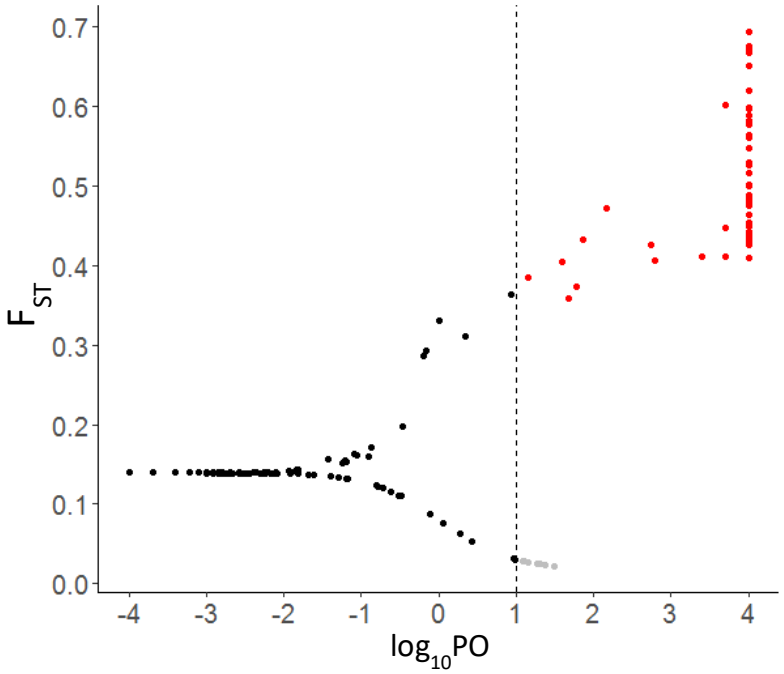


Figure 3.9. Results of Bayesian outlier analyses for 1,759 SNPs using *Bayescan*. Each dot represents a SNP with its F_{ST} plotted against $\log_{10}$PO (PO = posterior odds). SNPs with $\log_{10}$PO > 1 are identified as outliers. SNPs with $\log_{10}$PO = 1000 and -1000 are shown as $\log_{10}$PO = 4 and -4 respectively for clearer visualisation. Red and grey dots represent markers under divergent selection and balancing selection respectively.

A total of 76 SNPs were identified as outliers as they were significant in both outlier analyses. Distribution of these markers' maximum change in allele frequency, cline centre, cline width and cline slope are shown in Table 3.6 and Figure 3.10. Cline centre positions showed a peak at 1,900 – 2,000 m (Figure 3.10b). Cline width showed some variability (Figure 3.10c), while cline slope (< 0.01) were small for most markers (Figure 3.10d). Clines centred at around 1,900 m showed especially steep slopes (Figure 3.10e). However, cline width did not show correlation with cline centre position (Figure 3.11). The *STRUCTURE* cline, which acts as the "average" cline for tests of coincidence and concordance, has a centre and width of 1,852 and 866 m respectively. Among the 76 outlier clines, 51 and 40 were coincident and concordant with the *STRUCTURE* cline respectively; while 18 were both coincident and concordant with the *STRUCTURE* cline (Table 3.6**,** Figure 3.11). These indicate that clines show significantly variable coincidence and concordance. In general, clines with more similar cline centres are more likely to be coincident. The leaf dissection cline is coincident but discordant with the *STRUCTURE* cline.

Likelihood profiles for selected markers covering the observed range of cline centres and widths can be found in Figure 3.12.

3.5.5 Potential association between outlier markers and functional genes

Blast searches of outlier markers against scaffolds of unpublished *Senecio squalidus* draft genome and the non-redundant protein sequences database of NCBI led to the identification of 17 genes located within or in proximity (within 10 kb) of the markers (see Figure 3.7 for clines; Table 3.7 for position of markers on scaffolds or distance from functional genes) associated with different functions. Some outlier genes' functions have no apparent association with adaptation to contrasting environments on Mount Etna; while other genes identified are involved in photosynthesis, defence response, photoperiodism, flowering, metal ion binding, and response to UV, which could underlie environmental or ecological adaptation, or reproductive isolation. The details about these genes, including blast results, outlier analyses results, cline statistics, gene annotations, and information about the markers nearest to these genes are summarised in Table 3.7.

Table 3.6. Cline centres and widths of the 76 outlier nextRAD markers, and results of likelihood ratio tests of coincidence (centre) and concordance (width). Markers are arranged in the order of cline centre. Comparisons in likelihood ratio tests differed by one degree of freedom and significance was Bonferroni corrected for multiple testing. † in column one indicates that marker is used in estimating effective selection against hybrids. * in column four and five indicates significance hence non-coincidence or non-concordance with the *STRUCTURE* cline.

nextRAD marker ID	Cline centre (±2InL) (m)	Cline width (±2InL) (m)	2*Change in InL; centre fixed to *STRUCTURE* cline	2*Change in InL; width fixed to *STRUCTURE* cline
4163	1327 (1215-1420)	736 (550-995)	130.69*	1.17
10586	1329 (1200-1427)	572 (393-845)	135.00*	4.49
6372	1332 (1226-1431)	956 (749-1249)	99.96*	0.63
3207	1363 (1258-1498)	554 (328-971)	25.25*	2.23
21407	1381 (1245-1508)	1041 (758-1417)	55.23*	1.38
10167	1397 (1291-1493)	1067 (856-1352)	80.02*	3.51
10535	1447 (1314-1564)	1493 (1184-1943)	44.18*	24.76*
8777†	1463 (1401-1830)	80 (0-997)	4.56	1.85
7906	1472 (1393-1550)	763 (598-984)	69.19*	1.01
14794	1473 (1329-1595)	1258 (977-1680)	39.93*	8.84
12475	1491 (1398-1579)	1021 (820-1302)	57.35*	2.20
13348	1500 (1387-1631)	154 (2-462)	29.75*	8.98
9606†	1517 (1442-1725)	90 (0-955)	9.13	2.86
5376	1525 (1454-1596)	641 (505-824)	65.17*	5.74
11077	1526 (1436-1615)	905 (716-1163)	42.81*	0.13
4510	1531 (1440-1619)	1145 (935-1419)	46.66*	7.80
10491	1550 (1448-1653)	1030 (806-1349)	28.84*	1.98
14786	1558 (1458-1664)	991 (769-1302)	24.87*	1.10
4021	1568 (1494-1642)	826 (667-1035)	48.41*	0.18
6913	1584 (1447-1698)	692 (460-1004)	15.81*	1.46
185	1592 (1480-1718)	633 (308-1135)	10.98	1.16
13784	1620 (1543-1708)	484 (298-787)	17.14*	4.99
14946	1622 (1518-1787)	75 (0-821)	7.81	4.51
18343	1634 (1557-1712)	818 (661-1020)	28.05*	0.27
2237	1650 (1556-1744)	1010 (805-1286)	17.54*	1.74
7081	1672 (1605-1739)	753 (617-929)	27.34*	1.77
13459	1673 (1573-1769)	1085 (874-1375)	12.82	4.23
13363	1677 (1566-1789)	1165 (901-1542)	9.13	5.27
22657	1684 (1575-1795)	968 (750-1278)	8.27	0.71
10685	1722 (1589-1832)	827 (563-1179)	4.85	0.13
9343†	1748 (1407-1894)	821 (0-1141)	1.80	0.13
9311†	1767 (1637-1910)	659 (346-963)	1.50	1.96
2677†	1771 (1667-1880)	728 (436-1038)	1.97	0.69
2676	1792 (1713-1873)	644 (491-848)	2.16	4.63
4452	1884 (1855-1916)	64 (1-201)	4.45	18.46*
16721	1887 (1876-1924)	31 (2-146)	8.52	33.72*
10791	1892 (1862-1988)	51 (0-181)	5.09	21.28*

Table 3.6 (continued)

nextRAD marker ID	Cline centre (±2lnL) (m)	Cline width (±2lnL) (m)	2*Change in lnL; centre fixed to *STRUCTURE* cline	2*Change in lnL; width fixed to *STRUCTURE* cline
3443	1896 (1881-1937)	51 (3-173)	14.58*	45.76*
3992†	1901 (1805-1998)	1279 (1035-1623)	1.09	14.08*
4731†	1902 (1883-1988)	1 (0-102)	6.57	4.36
2952	1906 (1874-1952)	103 (4-299)	7.90	19.40*
2807†	1909 (1802-2015)	1270 (998-1658)	1.17	10.51
7971	1917 (1880-1989)	42 (0-159)	9.60	11.22
3508†	1922 (1830-2017)	1309 (1070-1627)	2.36	17.98*
1997†	1925 (1836-2017)	1269 (1042-1569)	2.77	16.06*
2180†	1929 (1842-2022)	1280 (1053-1592)	3.04	16.75*
3758†	1931 (1846-2019)	1237 (1020-1519)	3.44	14.45*
1092†	1932 (1843-2024)	1279 (1048-1586)	3.57	14.76*
3502†	1932 (1846-2020)	1236 (1022-1521)	3.24	16.43*
492†	1932 (1846-2023)	1243 (1027-1531)	3.44	14.45*
2052†	1933 (1793-2074)	1590 (1224-2166)	3.35	16.06*
143†	1933 (1844-2023)	1266 (1044-1562)	3.57	14.76*
3468†	1933 (1844-2023)	1270 (1042-1568)	1.38	23.73*
3564†	1933 (1846-2025)	1262 (1040-1562)	3.57	14.77*
2025†	1933 (1847-2021)	1240 (1026-1531)	3.57	14.76*
3165†	1933 (1847-2022)	1242 (1026-1527)	3.57	14.77*
2369†	1933 (1848-2022)	1244 (1025-1531)	3.57	14.77*
2579†	1933 (1848-2022)	1241 (1025-1532)	3.41	16.04*
2671†	1933 (1848-2022)	1242 (1028-1525)	3.47	15.87*
1011†	1934 (1844-2026)	1244 (1021-1552)	3.42	13.66*
3540†	1934 (1845-2025)	1267 (1044-1566)	3.46	15.92*
4473†	1936 (1848-2027)	1254 (1034-1553)	3.86	15.19*
1758†	1936 (1849-2026)	1258 (1035-1555)	3.79	15.34*
1357†	1936 (1849-2027)	1253 (1033-1549)	3.78	15.39*
1020†	1936 (1850-2027)	1256 (1033-1548)	3.73	15.50*
740†	1936 (1850-2027)	1251 (1032-1541)	3.76	15.43*
2994†	1937 (1849-2026)	1256 (1032-1549)	3.73	15.50*
1640†	1937 (1849-2027)	1260 (1037-1557)	3.79	15.34*
2668†	1937 (1850-2027)	1255 (1032-1548)	3.79	15.34*
776†	1938 (1850-2029)	1272 (1046-1573)	3.79	16.27*
4140	1940 (1882-1989)	74 (0-234)	7.88	9.88
2101†	1943 (1851-2035)	1265 (1027-1585)	3.99	13.89*
625†	1952 (1813-2088)	1539 (1174-2097)	2.12	20.52*
14157†	1983 (1876-2101)	990 (768-1309)	5.96	1.08
15757†	2000 (1906-2103)	1147 (918-1471)	9.99	6.30
1509†	2025 (1931-2132)	817 (621-1093)	14.63*	0.17

Table 3.7. Information of each marker containing outlier SNPs and their respective blastx and GO search results.

nextRAD marker ID	S. squalidus genome scaffold ID (size)	nextRAD marker position on scaffold	Outlier analysis			Cline analysis				
			F_{ST}	adj. P-value	log_{10}(PO)	Cline centre (±2InL) (m)	Cline width (±2InL) (m)	Pmin (±2InL)	Pmax (±2InL)	Cline slope
185	68665 (2036763)	477445-477595	0.4800	0.0065	∞	1592 (1480-1718)	633 (308-1135)	0.04 (0.00-0.10)	0.83 (0.75-0.94)	0.0012
2952	67856 (19539)	13784-13934	0.7395	0.0000	∞	1906 (1874-1952)	103 (4-299)	0.05 (0.02-0.08)	0.93 (0.86-0.98)	0.0086
3564	52973 (572)	183-332	0.4998	0.0020	∞	1933 (1846-2025)	1262 (1040-1562)	0	1	0.0008
4021	56 (21438)	17270-17420	0.6637	0.0000	∞	1568 (1494-1642)	826 (667-1035)	0	1	0.0012
4510	68720 (2345901)	28652-28802	0.5279	0.0001	∞	1531 (1440-1619)	1145 (935-1419)	0	1	0.0009
4731	68209 (1351264)	594497-594647	0.6335	0.0000	∞	1902 (1883-1988)	0.59 (0-102)	0.09 (0.05-0.14)	0.88 (0.78-0.94)	1.3406
7081	5706 (65989)	34626-34776	0.6885	0.0000	∞	1672 (1605-1739)	753 (617-929)	0	1	0.0013
9343	4 (2125992)	1879622-1879772	0.5749	0.0000	∞	1748 (1407-1894)	821 (0-1141)	0.13 (0.00-0.03)	1.00 (0.55-0.98)	0.0011
10491	68326 (3141945)	2793768-2793918	0.5714	0.0000	∞	1550 (1448-1653)	1030 (806-1349)	0	1	0.0010
10586	734 (2210660)	1278345-1278446	0.6579	0.0000	∞	1329 (1200-1427)	572 (393-845)	0	1	0.0017
11077	68729 (4218231)	3370224-3370374	0.6038	0.0000	∞	1526 (1436-1615)	905 (716-1163)	0	1	0.0011
12475	67909 (194016)	137067-137215	0.5230	0.0024	∞	1491 (1398-1579)	1021 (820-1302)	0	1	0.0010
13459	67888 (10451821)	9651167-9651319	0.5857	0.0000	∞	1673 (1573-1769)	1085 (874-1375)	0	1	0.0009
14157	68656 (668629)	331438-331588	0.6995	0.0000	∞	1983 (1876-2101)	990 (768-1309)	0	1	0.0010
15757	67999 (7338612)	3875968-3876118	0.4962	0.0069	∞	2000 (1906-2103)	1147 (918-1471)	0	1	0.0009
16721	68558 (99546)	34967-35117	0.8587	0.0000	∞	1887 (1876-1924)	31 (2-146)	0.02 (0.00-0.06)	0.96 (0.92-0.99)	0.0301
22657	68794 (55174)	8239-8389	0.5670	0.0000	∞	1684 (1575-1795)	968 (750-1278)	0	1	0.0010

Table 3.7 (continued)

nextRAD marker ID	Blastx and GO search	
	Top hit	**Blast top hit accession**
185	traB domain-containing protein isoform X1 [Lactuca sativa]	XP_023755563.1
2952	WD repeat-containing protein 62 isoform X1 [Helianthus annuus]	XP_022034050.1
3564	photosystem II CP47 chlorophyll apoprotein [Schoenolirion croceum]	YP_009335530.1
4021	uncharacterized protein LOC110931622 [Helianthus annuus]	XP_022030699.1
4510	putative disease resistance protein RGA3 [Lactuca sativa]	XP_023751920.1
4731	reverse transcriptase domain-containing protein [Artemisia annua]	PWA89409.1
7081	uncharacterized protein LOC111897622 [Lactuca sativa]	XP_023749357.1
9343	cellulose synthase-like protein G3 [Cynara cardunculus var. scolymus]	XP_024996549.1
10491	hypothetical protein CTI12_AA337030 [Artemisia annua]	PWA65181.1
10586	GIGANTEA-like protein [Cynara cardunculus var. scolymus]	KVI07235.1
11077	F-box/FBD/LRR-repeat protein At1g13570-like [Helianthus annuus]	XP_022020143.1
12475	Concanavalin A-like lectin/glucanase, subgroup [Cynara cardunculus var. scolymus]	KVI02890.1
13459	aspartokinase 2, chloroplastic-like isoform X2 [Lactuca sativa]	XP_023768428.1
14157	uncharacterized protein LOC111887355 [Lactuca sativa]	XP_023739289.1
15757	HCP-like superfamily protein [Artemisia annua]	PWA96304.1
16721	uncharacterized protein LOC110866042 isoform X2 [Helianthus annuus]	XP_021971094.1
22657	RHOMBOID-like protein 14 [Artemisia annua]	PWA48927.1

Table 3.7 (continued)

nextRAD marker ID	Blastx and GO search GO terms
185	C: part of membrane, integral component of membrane
2952	/
3564	F: enables chlorophyll binding; P: involved in photosynthetic electron transport chain, photosynthesis, protein-chromophore linkage, photosynthesis, light reaction; C: part of photosystem, membrane, integral component of membrane
4021	/
4510	F: enables nucleotide binding, ATP binding, ADP binding; P: involved in defence response
4731	C: part of plastid
7081	/
9343	F: enables transferase activity, transferase activity, transferring glycosyl groups, cellulose synthase activity, cellulose synthase (UDP-forming activity); P: involved in plant-type primary cell wall biogenesis, celulose biosynthetic process, cell wall organisation; C: part of Golgi membrane, Golgi aparatus, trans-Golgi network, plasma membrane, membrane, integral component of membrane
10491	/
10586	P: involved in regulation of photoperiodism, flowering
11077	C: part of membrane, integral component of membrane
12475	F: protein kinase activity, ATP binding, carbohydrate binding; P: protein phosphorylation; C: part of membrane, integral component of membrane
13459	F: enables aspartate kinase activity, kinase activity; P: cellular amino acid biosynthetic process, threonine biosynthetic process, lysine biosynthetic process via diaminopimelate, phosphorylation
14157	/
15757	C: part of membrane, integral component of membrane
16721	/
22657	F: enables endopeptidase, serine-type endopeptidase, peptidase activity, hydrolase activity, metal ion binding; P: involved in proteolysis, positive regulation of protein processing, membrane protein intracellular domain proteolysis, cellular response to unfolded protein, cellular response to UV, post-translational protein modification, positive regulation of secretion, membrane protein proteolysis involved in retrograde protein transport, ER to cytosol; C: part of mitochondrion, plasma membrane, membrane, integral component of membrane, integral component of endoplasmic reticulum membrane, mitochondrial membrane, endoplasmic reticlumn quality control compartment

Table 3.7 (continued)

nextRAD marker ID	Remarks	Nearest non-outlier SNP 1		Nearest non-outlier SNP 2	
		Distance (bp)	F_{ST}	Distance (bp)	F_{ST}
185	/	393373	0.1916	/	/
2952	/	/	/	/	/
3564	/	/	/	/	/
4021	/	/	/	/	/
4510	/	130264	0.1260	/	/
4731	/	89765	0.1930	117220	0.0039
7081	/	/	/	/	/
9343	/	1011543	0.0340	102073	0.2100
10491	The gene is around 8050bp away from marker.	1327091	0.1459	/	/
10586	/	147116	0.0780	50842	0.0470
11077	This gene appears in the list of differentially expressed genes in Chapman *et al.* (2009).	368743	0.0580	556297	0.0000
12475	/	/	/	/	/
13459	/	144246	0.0800	/	/
14157	/	/	/	/	/
15757	/	350097	0.0000	203419	0.1160
16721	/	/	/	/	/
22657	/	/	/	/	/

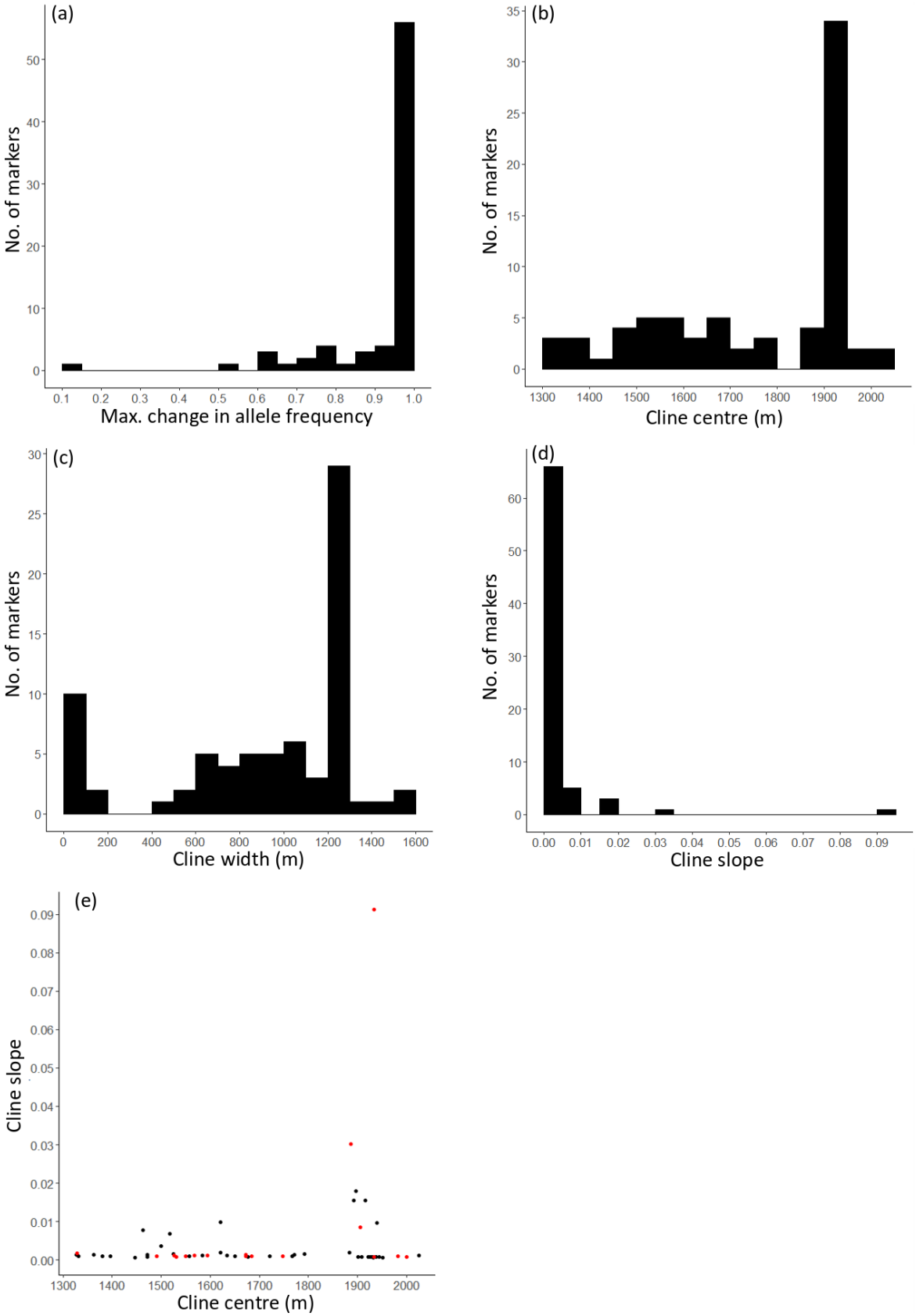


Figure 3.10. Distributions of different cline parameters in nextRAD outlier SNPs. (a) Maximum change in allele frequencies. (b) Distribution of cline centres. (c) Distribution of cline widths. (d) Distribution of cline slopes. (e) Cline centres and their corresponding cline slope of each SNP. Red dots represent the 17 outlier markers that are gene-annotated.

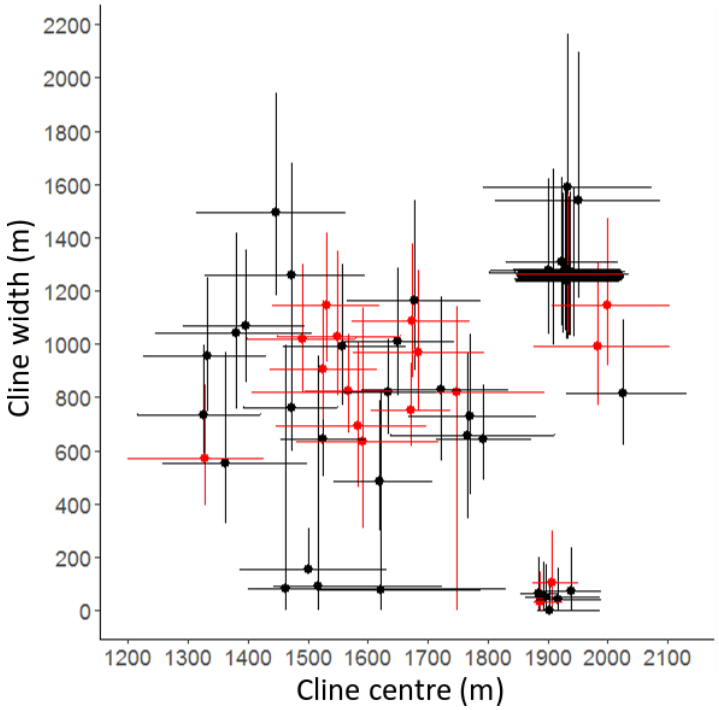


Figure 3.11. Variable cline coincidence and concordance of outlier clines. Maximum likelihood-estimated cline widths and cline centres of each outlier SNP are plotted as dots, while support limits are plotted as lines extending from the dots. Black and red data points represent nonoutlier and outlier SNPs, respectively. Refer to Table 3.6 for test results.

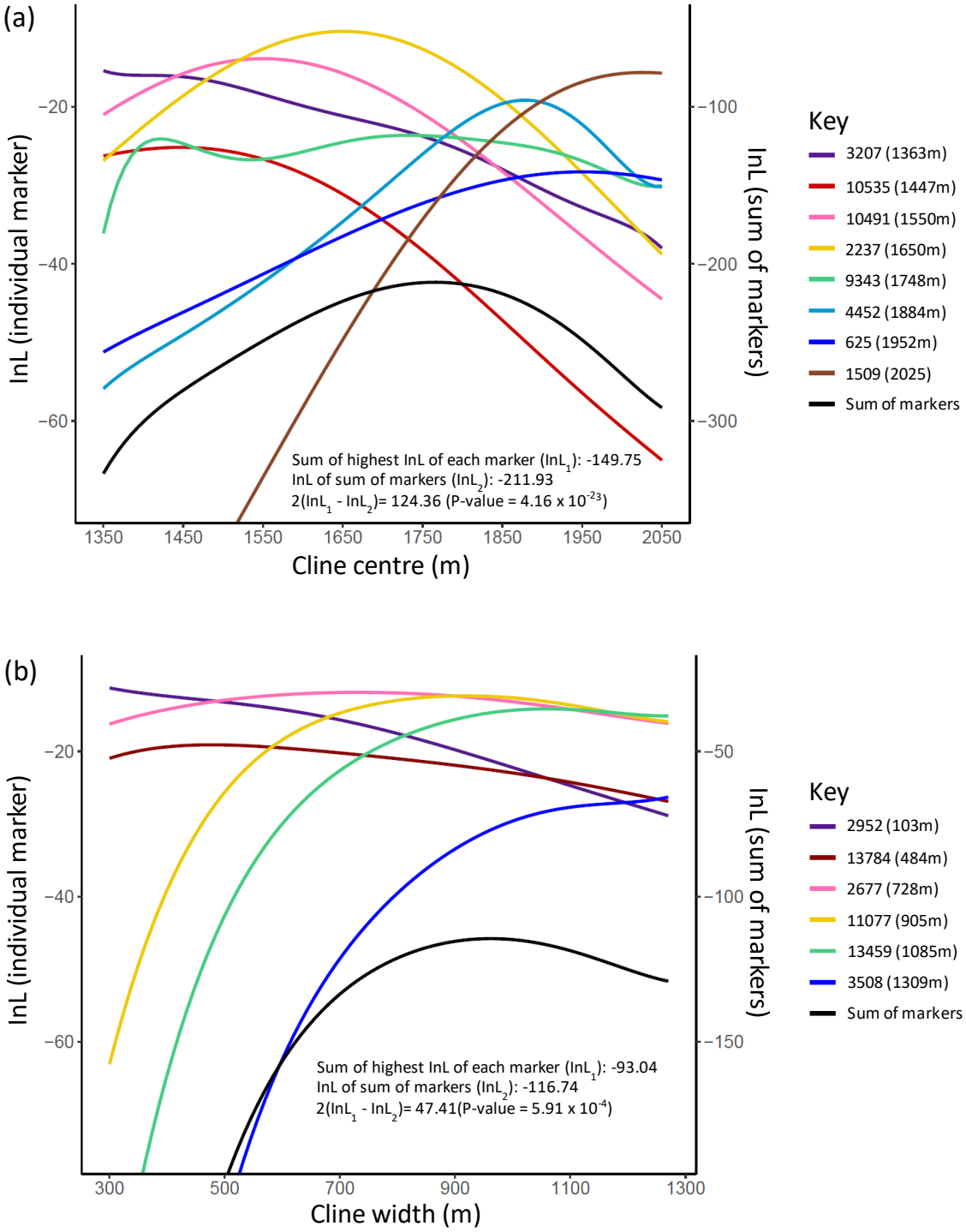


Figure 3.12. Likelihood profiles for selected markers that have a range of maximum likelihood centres and widths (coloured curves), and sum of markers (black curve). Under the keys, numbers represent marker IDs and numbers in brackets are the markers' respective maximum likelihood cline centre or width. Likelihood ratio tests (details in graphs) reveal that clines are significantly non-coincident and non-concordant.

3.5.6 Strength of selection

Population linkage disequilibrium of the 40-marker data set with coincident and concordant clines is elevated at the hybrid zone centre at around 1,900 – 2,000 m in elevation (Figure 3.13). Linkage disequilibrium estimated for the cline centre is 0.3915 (95% CI = 0.3681 – 0.4119). The dispersal rate was estimated to be 0.56 – 1.27 km/gen (corresponding to 144–323 m in elevation) and the corresponding average effective selection is 0.1565 (95% CI = 0.1472 – 0.1648) and 0.7831 (95% CI = 0.7521 – 0.8144), assuming recombination rate of 0.1 and 0.5, respectively.

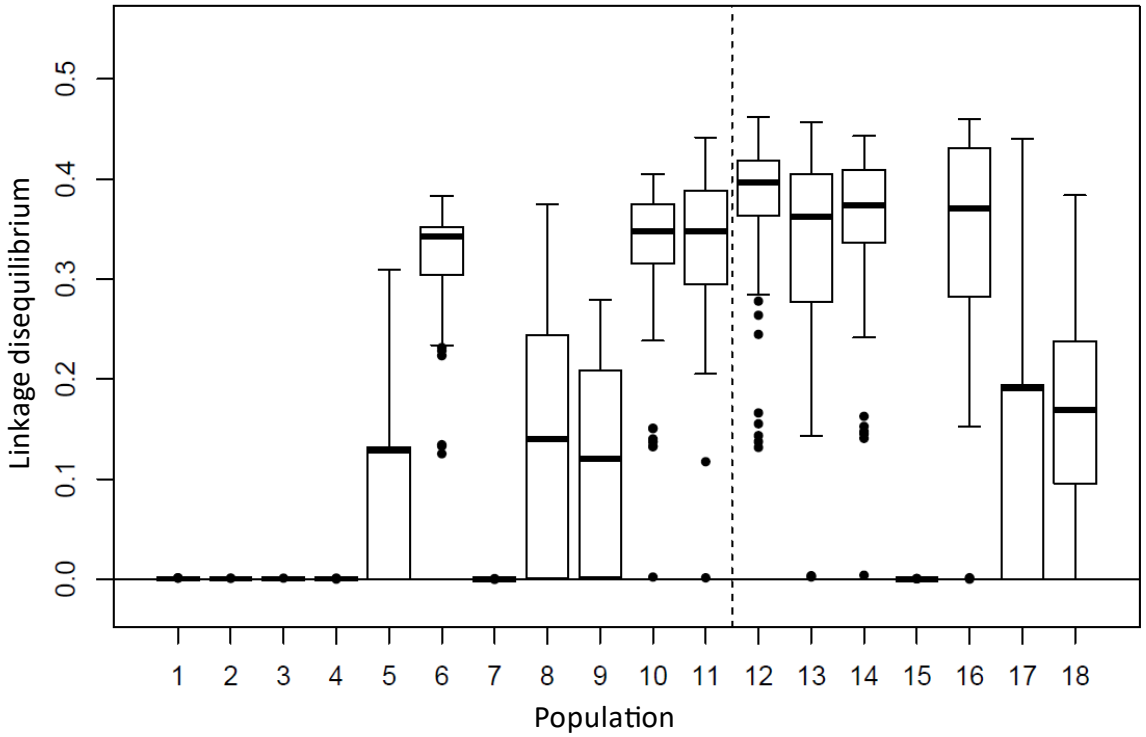


Figure 3.13. Linkage disequilibrium obtained using the 40 coincident and concordant clines used in estimating selection and dispersal. The vertical dotted line is the approximate inferred position of the hybrid zone centre (1,900 – 2,000 m in altitude).

3.5.7 How wide are the genomic regions affected by divergent selection?

Identifying non-outlier SNPs (noSNPs) that are closest to the 17 outliers (oSNPs) through blasting of adjacent regions (same criteria as identifying genomics scaffolds for outliers), it was revealed that the distance between the noSNPs and oSNPs (that fall within the same scaffold) range from around 50 kb to 1.3 Mb, with F_{ST} for noSNPs ranging from 0 to 0.19. For some genomic regions, noSNPs with $F_{ST} < 0.047$ are located as close as 51 kb to oSNPs (e.g., marker 10586 in Table 3.7); while in other regions, the distance was much longer and the noSNP's F_{ST} was not as low (e.g., 400 kb and $F_{ST} = 0.20$ for marker 185 in Table 3.7). This indicates that the signal of selection in the genomes of Etnean *Senecio* species may be very localised. Further work has to be done to study the selection patterns (and evaluate the effect of missing data) in detail. For instance, denser sequencing would allow more markers to be identified in proximity to oSNPs to examine changes in their and noSNPs' levels of polymorphisms more precisely.

3.6 Discussion

Hybrid zones are a powerful tool to study the dynamics of phenotypic and genotypic divergence between closely related species; also, recently diverged species offer great opportunities to catch "speciation in action" by studying the patterns of gene flow and genomic heterogeneity in taxa subject to divergent selection (Via, 2009). *S. aethnensis* and *S. chrysanthemifolius* on Mount Etna are characterised by a hybrid zone and recent divergence less than 200,000 years ago. Using genome-wide SNPs in samples comprising both species and their hybrids, I investigated demographic features, patterns of gene flow and selection in this system, and factors that are probably contributing to maintenance of divergence in the face of ongoing gene flow between the two species.

3.6.1 Demographic history of speciation in Etnean *Senecio*

Various demographic models implemented in *dadi* (see Figure 3.2 for visualisation of models) were used to reconstruct the most likely speciation scenario and demographic features in the divergence of *S. aethnensis* and *S. chrysanthemifolius* on Mount Etna.

Our analysis revealed that the *IM2M* model fits the data best. This model has not been considered in previous studies in *Senecio*, but some of the parameter estimates are similar to that obtained previously. In particular, migration between the two pure species inferred here is similar to the estimate from the most likely model in Filatov *et al.* (2016), which is less than one migrant per generation on average (average between M_A and M_B; number of migrants per generation is higher in other geographically closer comparisons in the current study). Given models with (*SC_IM2M*) and without (*IM2M*) secondary contact both have high relative likelihoods, there is little power in the data to distinguish between these demographic scenarios; it may also be seen as a lack of evidence for the secondary contact scenario as the more complex *SC_IM2M* model did not significantly improve fit to data (consistent with Filatov *et al.*, 2016). What is common between both models is that they both involve heterogeneous gene flow. This indicates that a fraction of the genome [42% (33 - 52)%] for *IM2M* and 74% (68 – 80%) for *SC_IM2M*; parameter P in Table 3.5] introgresses considerably slower compared to the rest of the genome, possibly due to intrinsic or extrinsic selection against introgressed alleles in these regions. Many studies have demonstrated heterogeneous divergence in

different genomes and that divergent selection has a crucial role in genomic heterogeneity (reviewed in Nosil *et al.*, 2009a).

Another interesting point to note is that gene flow is highly asymmetric, as revealed by four models for the pure species comparison and *IM* analyses of other pure species/ hybrid comparisons. Gene flow from *S. chrysanthemifolius* to *S. aethnensis* is 1.24 to 1.70 times higher than that from *S. aethnensis* to *S. chrysanthemifolius*; while gene flow is always (1.6 to 3 times) greater from pure species to hybrids, compared to gene flow from hybrids to pure species. This low rate of introgression from hybrid to pure species could be caused by various reasons, such as lower reproductive fitness in hybrids, and low rate of backcrossing. The directionality of gene flow helps to explain the apparent contradiction between plentiful evidence of hybridisation at intermediate elevations and relatively modest estimates of gene flow between the pure species at the two extremes of the elevational cline.

3.6.2 Variable coincidence and concordance among phenotypic and genotypic clines

I fitted clines for leaf dissection and genetic markers using the same individuals. Likelihood ratio tests revealed that clines have significantly variable coincidence and concordance. Although clines studied here have variable coincidence, they are all centred in the putative hybrid zone (1,200 – 2,000 m in elevation), with only a few exceptions that are centred above 2,000 m (this corresponds to the observation in the *STRUCTURE* analysis where there is a little admixture in high-elevation *S. aethnensis*). These environmental and ecological factors probably vary in different ways with regard to elevations within the hybrid zone. Other studies, such as those on the marine snails *Littorina saxatilis*, have also suggested displaced clines among different environmental transitions that would lead to selection acting on different loci independently (e.g., Hollander, Galindo, & Butlin, 2015). Like the current study, many studies on these snails have also identified genomic regions under divergent selection and associated with adaptive phenotypic traits (Hollander *et al.*, 2015; Le Pennec *et al.*, 2017; Westram *et al.*, 2014).

Analysing allozymes, simple sequence repeats (SSR) and various groups of phenotypic measurements, Brennan *et al.* (2009) found highly coincident and concordant clines among all of them, in which they are centred at 6.67 – 7.82 km (corresponding to elevations somewhere between 1,515 – 1,928 m) with widths of 1.49 – 3.7 km. In particular, they found the leaf *STRUCTURE* cline to be centred at 6.67 (6.25 – 6.97) km with width of 1.49 (0.06 – 2.39) km. These support limits translate into cline centre within the elevational range of 1,515 – 1,795 m and width of more than 400 m in elevation. In the study by Brennan *et al.* (2009) leaf *STRUCTURE* cline does not seem to be coincident with the leaf dissection cline in this study, while the cline width in this study is well within the range reported by Brennan *et al.* (2009) [in the current study, cline centre = 1,910 (1,882 – 1,935) m; width = 127 (51 – 196) m]. The discrepancy could be due to: (a) more than one leaf traits were combined in the leaf *STRUCTURE* cline in Brennan *et al.* (2009). Individual leaf trait clines might have centres and widths different from leaf dissection; (b) samples in Brennan *et al.* (2009) were collected from a few areas located on different sides of Mount Etna, while samples in this study were collected along a more or less straight transect on the southern side of the mountain. Hence, results in the two studies should be compared with caution.

3.6.3 Strong selection against hybrids on Mount Etna

Effective selection against hybrids inferred in this study (0.16–0.78 assuming recombination between the markers is $r = 0.1 – 0.5$) is much higher than that in Brennan *et al.* (2009) (0.02 – 0.11 for $r = 0.1 – 0.5$, respectively) for the same species. Dispersal rate inferred in this study (0.56 – 1.27 km/gen for $r = 0.1 – 0.5$) is also greater than that in Brennan *et al.* (2009) (0.17 – 0.38 km/gen for $r = 0.1 – 0.5$). Both studies used only coincident and concordant clines in estimating selection, thus the difference could be related to the type of markers used (allozymes and microsatellites in Brennan *et al.* (2009); SNPs in the current study). For instance, microsatellites have multiple alleles, hence higher heterozygosities and mutation rate than SNPs (Mueller, 2004); these properties could lead to different estimates of linkage disequilibrium and subsequent estimates. In our analysis, linkage disequilibrium is elevated at the hybrid zone centre, which is expected for a hybrid zone experiencing strong selection. However, linkage disequilibrium in the high-elevation localities for *S. aethnensis* is

also quite high. This could be caused by an influx of alleles from *S. chrysanthemifolius*, as revealed by the presence of admixed individuals in high-elevation localities in the *STRUCTURE* analysis, and asymmetric migration in demographic modelling. The asymmetry in gene flow could be deal to *S. chrysanthemifolius* being more opportunistic that *S. aethnensis* (personal communications with collaborators), hence seeds of the former have a higher chance of germinating and surviving when the right conditions arise at high elevations. Another reason could be that breezes from the ocean to land would help carry *S. chrysanthemifolius* seeds to high elevations.

Estimates of selection and dispersal from a recombination rate of 0.5 (the higher values) are probably closer to reality as it is unlikely that all markers used are tightly clustered. Our selection estimate from this recombination rate corresponds to a 78% drop in fitness in hybrids in the centre of hybrid zone, which indicates that the *Senecio* cline is maintained by fairly strong selection. This is one of the highest estimates reported (Table 3.1). Previous studies on salamanders (Alexandrino *et al.*, 2005) and skinks (Phillips *et al.*, 2004) have reported values similar to the current study, yet they calculated selection using $s = 8\sigma^2/w^2$ while our study used $s^* = 4\sigma^2/w^2$ (Table 3.1; $s^* = 4\sigma^2/w^2$ measures the difference in mean fitness between centre and edge of hybrid zone, and measuring this difference in mean fitness makes selection estimates comparable between different forms of selection; whereas $s = 8\sigma^2/w^2$ assumes half of the individuals are heterozygotes at the hybrid zone centre, where $s = 2s^*$). In other words, the estimate in the current study (0.78) is at least two, and up to more than 200, times larger than that in other recent studies presented in Table 3.1 after accounting for the fold differences in the formulas used. It is worth noting that Brennan *et al.* (2009) also calculated effective selection in Etnean *Senecio* with $8\sigma^2/w^2$, though this does not account for the difference between the estimates in their and the current studies either.

The inferred dispersal (0.56 – 1.27 km/gen) is rather small, compared to what one would expected for wind-dispersed herbaceous plants. This, together with strong selection against hybrids and demographic modelling showing around one migrant per generation, collectively indicate that gene flow is restricted, probably due to divergent selection. Selection against hybrids has only been shown to manifest as hybrid breakdown in F_2 plants in two crosses in three studies (Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009), thus larger-scale studies are required to further investigate the effect on hybrid establishment and fitness, which is carried out in *Chapter 5*.

Multifarious selection is evident in this system, as clines of many outlier markers are relatively wide (hence selection at many individual traits is unlikely to be strong) while a few show rather narrow clines with widths under 100 m; and only a small proportion of clines is coincident and concordant. The strong selection against hybrids and occurrence of multifarious selection probably facilitates local adaptation in the two species and their divergence. Future work that analyses how potential environmental or ecological selective forces (for example, whether they change gradually or abruptly along the gradient and how different their effects are) had shaped the system would be particularly fascinating for this young but complex system.

3.6.4 Intrinsic and extrinsic causes in the divergence of Etnean *Senecio*

Regardless of the speciation scenario at the start of divergence, Etnean *Senecio* has experienced heterogeneous gene flow. One question that arises from this is whether selective sweeps [leading to (near-) fixation of SNPs in populations of pure species] or barriers to gene flow (selection in the hybrid zone in this case) cause divergence in this differentiated portion of genome (Tavares *et al.*, 2018). Both of these would show signatures of divergent selection, and disentangling between the two would require higher-resolution genomic data. It would also be interesting to quantify the regions and the extent of frequent or rare migration in the genome in future studies.

Research on intrinsic incompatibilities is often disconnected from studies of extrinsic selection, making integrating different components of reproductive isolation challenging (Seehausen *et al.*, 2014). With both intrinsic incompatibility and multifarious divergent selection, the Etnean *Senecio* allows one to combine both directions of research to study the interplay between intrinsic and extrinsic processes. This adds to the limited literature in other systems which also have evidence for both, including in killfish, *Lucania*, where there is extrinsic isolation caused by salinity and intrinsic incompatibilities (Fuller, 2008); and in the copepod *Tigriopus californicus*, where interpopulation hybrids show adaptation to different thermal conditions and intrinsic incompatibilities (Ellison & Burton, 2006, 2008; Willett, 2010). However, unlike these systems, the Etnean *Senecio* is parapatric and multifarious selection might be crucial in creating strong enough

reproductive isolation to allow the accumulation and maintenance of genetic divergence in the system (Seehausen *et al.*, 2014).

3.6.5 Maintenance of divergence despite gene flow

It is intriguing how the pair of sister *Senecio* species maintain their divergence while hybridising in the relatively small area of Mount Etna. Yet, this phenomenon is not uncommon - *Quercus* (oaks) and *Populus* (poplars) are well-studied groups with a "porous" species boundary. Studies in these two groups have shown that their species identities are maintained by an array of mechanisms, some of which have been presented or observed in the Etnean *Senecio*. These include, in oaks, divergent selection (Ortego, Gugger, & Sork, 2017; Scotti-Saintagne *et al.*, 2004), and asynchrony in flowering leading to assortative mating (Gailing & Curtu, 2014); in poplars, intrinsic incompatibility (Christe *et al.*, 2017; Roe *et al.*, 2014), and selection against hybrids (Christe *et al.*, 2016). It is plausible that a combination of divergent selection, strong cumulative selection against hybrids and difference in phenology (such as flowering times) helps to reinforce the reproductive isolation in the system. Given the prevalence of hybridisation between closely-related plant species across different families, these systems point to the possibility that more unstudied plant groups and species complexes might have extensive hybridisation while maintaining their respective species boundaries or divergence at sub-species levels. Studying these groups would greatly benefit the field of plant hybridisation, speciation with gene flow, and study of the role of gene flow, selection and introgression in adaptation and speciation.

In conclusion, I have shown that only a small proportion of differentiated loci (76 out of 1,769 studied) and strong cumulative multifarious selection at a handful of genes are probably involved in keeping the two Etnean *Senecio* species distinct. Although the exact mechanism of speciation in this system is far from clear, this study has added to the body of evidence that both intrinsic and extrinsic processes have roles in speciation and maintenance of divergence in the face of gene flow (Ellison & Burton, 2006, 2008; Fuller, 2008; Willett, 2010). To summarise, I have (a) shown heterogeneous, bidirectional gene flow, and population size changes in the past are characteristics of the speciation of Etnean *Senecio*; (b) discovered promising candidate genes that are potentially linked to local

adaptation; (c) inferred the likely presence of multifarious selective forces of different strengths; and (d) estimated strong selection against hybrids. It is crucial to extend this work by identifying the selective forces underlying divergence and linking them to elevation or ecology, analysing fine-scale genomic patterns that harbour regions of interests, the interplay between intrinsic incompatibilities and extrinsic, multifarious environmental selection in shaping the system, and the potential role of other genetic phenomena (such as genetic hitchhiking and chromosomal rearrangement) in causing divergence.

4. Genome scans reveal vast number of regions harbouring putative selective sweeps and genes related to adaptation in high- and low-elevation *Senecio* species on Mount Etna

4.1 Abstract

Adaptation is a crucial process in which species become fit in their environments. It is a consequence of natural selection over generations. Undergoing selection and adaptation would leave footprints in the genome known as selective sweeps - when a selective agent exerts selective pressure on a species, the advantageous mutations will spread in the population and eventually become fixed, resulting in the loss of diversity. Selective sweeps have been studied in many species adapted to different environmental conditions to identify the genes, functions and processes that were involved in adaptation, and also the effect of sweeps on nucleotide diversity. This can be done by examining the genomic patterns to look for signatures of sweeps, including changes in polymorphisms, site frequency spectra and linkage disequilibrium. In this study, the linkage map for Etnean *Senecio* was integrated with the *Senecio squalidus* draft genome. Relationships between recombination rate and nucleotide diversity was investigated using markers from previous studies. Moreover, using genomic sequences of *S. aethnensis* and *S. chrysanthemifolius*, this study reported the first genomic scans of selective sweeps and genome-wide patterns of polymorphisms. Consistent with theory, recombination rate and nucleotide diversity correlated positively and significantly in most of the linkage groups. Putative selective sweeps showed various strengths and age, and strong signals were revealed on all linkage groups. These regions that show strong sweep signals harbour many genes that are related to adaptation to contrasting environments on Mount Etna, such as those related to metal ions, defence and floral developments, some of which were not reported before. These genes create opportunities in the future to study plant-biotic and -abiotic interactions. More effort should also be put in creating a better draft genome and its annotations to allow more in-depth genomic studies in the system.

4.2 Introduction

In population genetics, one fundamental question is what shapes the patterns of genetic diversity within and among different species. For many years, population geneticists have been trying to disentangle the complex relationships among genetic polymorphisms, recombination and selection. Recombination is usually supressed near centromeres, and its rate varies among species, populations, sexes, individuals, and regions of the chromosome (e.g. Carneiro, Ferrand, & Nachman, 2009; Fernandes, Wlodzimierz, & Henderson, 2019; Nachman, 2002). Selection is the process in which certain traits and their underlying genes are accumulated or removed through time (e.g. Kimura 1961; 1968). Both recombination and selection can have far-reaching effects on the patterns of polymorphism in the genome (e.g. Aquadro, Begun, & Kindahl, 1994; Nachman, 2001; Roselius, Stephan, & Städler, 2005; Wang, Street, Scofield, & Ingvarsson, 2016).

There are two leading models of selection that explains the decrease in polymorphisms at linked genetic sites. The first model, the genetic hitchhiking model, depicts the scenario when adaptive mutations and their linked, neutral variations, sweep across populations to fixation (Smith & Haigh, 1974). In these regions of selective sweeps where polymorphism is low, the size of the regions affected by sweeps depends on the recombination rate and selective coefficient; whereas the strength of sweeps depends on the latter. The higher the selective coefficients (faster spread of mutations) and the lower the recombination rate, the bigger the region affected by sweeps is. If adaptative mutations are common, the genetic hitchhiking effect would lead to a positive correlation between recombination rate and the level of polymorphism (Begun & Aquadro, 1992; Kaplan, Hudson, & Langley, 1989). The second model, the background selection model, depicts the scenario when selection eliminates deleterious mutations and their linked neutral variations (Charlesworth, Morgan, & Charlesworth, 1993). Theories suggest that if the rate of deleterious mutations is high and selection is weak, this model would result in a positive correlation between recombination rate and the level of polymorphism (Hudson & Kaplan, 1995), unless there are also frequent selective sweeps by genetic hitchhiking (Wiehe, 1998). The two models are not mutually exclusive, as both positive selection that drives the fixation of adaptive mutations and negative selection that leads to the elimination of deleterious mutations (background selection) could work together to result in the decrease in the

level of heterozygosity at linked regions – the question then becomes how much effect each of them has on the observed patterns of polymorphisms (Nachman, 2001). However, these two models do not explain all the phenomena observed in low-recombining regions (with sweeps). For instance, demographic processes such as bottlenecks could increase polymorphism across loci between groups, as shown in *Arabidopsis lyrata*, where only some of the populations studied had low polymorphism in low recombining regions (Wright *et al.*, 2006).

A hard selective sweep (hereafter 'sweep') would leave three types of signals in the genome: (a) a reduction in the level of polymorphism locally (Smith & Haigh, 1974); (b) a shift in site frequency spectrum (SFS) towards high and low-frequency derived variants (Braverman, Hudson, Kaplan, Langley, & Stephan, 1995), as neutral variants linked and not linked to the beneficial variants increase and decrease in frequencies respectively (Braverman *et al.*, 1995; Fay & Wu, 2000); and (c) a pattern of linkage disequilibrium (LD) that is localised – high LD on each side of the beneficial variant and low LD between the loci on opposite sides of the beneficial variant (Kim & Nielsen, 2004; Stephan, Song, & Langley, 2006). To examine the pattern of polymorphisms, neutrality tests using statistics such as Tajima's D can be used, in which it would be negative in proximity to sweeps (e.g. Bachtrog, 2004; Orengo & Aguadé, 2004; Sabeti *et al.*, 2006). To detect shifts in SFS, several software tools have been developed based on the composite-likelihood-ratio (CLR) test (Nielsen *et al.*, 2005) to detect recent sweeps [such as *SweeD* (Pavlidis, Živković, Stamatakis, & Alachiotis, 2013) and *SweepFinder2* (DeGiorgio, Huber, Hubisz, Hellmann, & Nielsen, 2016)]. Nonetheless, given background selection could have a considerable effect in reducing genetic diversity in low-recombining regions (e.g. Akashi, Osada, & Ohta, 2012; Charlesworth, 2012; Charlesworth, Charlesworth, & Morgan, 1995), it could show similar patterns as sweeps. To alleviate this problem, McVicker, Gordon, Davis, & Green (2009) proposed the estimation of B-values, which are in proportion to local decline in genetic diversity or effective population size caused by background selection. This is implemented in *SweepFinder2* so that the effect of background selection is also accounted for alongside local recombination events (DeGiorgio *et al.*, 2016). Last but not least, Kim & Nielsen (2004) developed the ω statistic and a maximum likelihood framework that utilises LD to detect sweeps, which are implemented in software tools like *OmegaPlus* (Alachiotis, Stamatakis, & Pavlidus, 2012). Apart from searching for sweep signals with either LD or SFS, software tools such as *RAiSD* (Alachiotis & Pavlidis, 2018), were designed to look

for all three signatures of sweeps. *RAiSD* utilises the μ statistic, which is a composite evaluation test that tracts fluctuations in the amount of polymorphism, level of LD and SFS along chromosomes (Alachiotis & Pavlidis, 2018). Studying sweeps is extremely informative for understanding the evolutionary processes, as it is a major footprint of adaptation (Alachiotis & Pavlidis, 2018). Many studies have investigated sweeps using the aforementioned software tools in different species in order to infer what evolutionary events have shaped their similarities and differences. These include those in crop species or wild relatives of crops (e.g. wild populations of Atlantic salmon, *Salmo salar*: Lehnert *et al.*, 2019; African and Asian rice, and their wild progenitor, *Oryza sp.*: Ndjiondjop *et al.*, 2019), and non-model organisms (e.g. high and low elevation populations of Cushion Willow, *Salix brachista*: Chen *et al.*, 2019).

Although many tools are available to analyse selective sweeps, distinguishing between false-positive selective sweeps from real ones could be extremely challenging (Pavlidis & Alachiotis, 2017). One reason is that software tools that identify selective sweeps are affected by demographic processes, such as population bottlenecks, as these processes often result in a high rate of coalescent events; which would in turn lead to SNP patterns that resemble those of selective sweeps (Pavlidis *et al.*, 2017). Pavlidis *et al.* (2017) also pointed out that the combined effects of selection and demographic processes are not clear, and that events like bottlenecks are unfortunately very common, making it extremely difficult to distinguish between the two. Apart from demographic processes, inaccuracies (even small ones) in genomic sequences could lead to false-positives too (Mallick, Gnerre, Muller, & Reich, 2009). Using re-sequenced data with increased coverage in chimpanzee genes previously identified to be under positive selection, Mallick *et al.* (2009) found substantial decrease in the number of genes (from 59 to six only) to be under significant positive selection. They also found more stringent filtering of genomic data dramatically reduced significant signals. Comparing methods based on summary statistics of polymorphisms (such as F_{ST}) and SFS, Thornton & Jensen (2007) showed the former had more power and resulted in fewer false-positive signals in detecting recent selective sweeps. One way to verify signals of selective sweeps is to analyse the patterns of polymorphisms in the corresponding regions. In regions under hard selective sweeps, nucleotide diversity (π) is expected to be low due to the fixation of mutations while surrounding regions' π

remains moderate to high, hence creating a V-shape valley (Kim & Stephan, 2002). However, false-positives due to high error rate in sequencing data still cannot be ruled out.

So far, selection studies on the high-elevation *S. aethnensis* and low-elevation *S. chrysanthemifolius* on Mount Etna have been focused on particular traits and genetic loci which are underlying divergent selection. Brennan *et al.* (2009) demonstrated clinal variations along the elevational gradient of Mount Etna in three quantitative trait groups (inflorescence structure, leaf structure, seed and fruit structure) which were likely under extrinsic environmental selection. They have also found relatively strong intrinsic selection against hybrids in allozymes and simple sequence repeat loci. In *Chapter 3* (Wong *et al.*, 2020), it has been shown that multiple selective agents are likely present on Mount Etna to maintain strong cumulative selection against hybrids [up to 0.78 (on a scale of 0 to 1, with 0 being no selection at all and 1 being fully selected against), one of the highest reported in literature among all organisms] and divergence between the two species. Various SNPs were underlying genes that were related to these selective agents such as UV and metal ions. However, no studies have been done to investigate the genome-wide patterns of selection in this system yet. With the availability of a draft reference genome of *S. squalidus* (the hybrid species between *S. aethnensis* and *S. chrysanthemifolius*) and genomic sequencing data of the two species, novel genomic studies for the Etnean *Senecio* system are now feasible. Using the linkage map of Chapman *et al.* (2013) and the *S. squalidus* draft genome, I integrated the linkage map with genomic sequences to investigate the relationship between nucleotide diversity and recombination rate. Using genomic sequences and software tools mentioned above, I looked at the distribution of polymorphisms on each linkage group and carried out a genome-wide scan for selective sweeps to look for regions that harbour putative targets of selective sweeps. I hypothesise that (a) recombination rate and nucleotide diversity would have a positive correlation, consistent with theory of genetic hitchhiking (Smith & Haigh, 1974); (b) genomic regions showing strong sweep signals would underlie traits and genes useful for adaptation to the contrasting environments on Mount Etna; (c) there would be multiple sweep signals on various linkage groups, with various strengths and ages, as multifarious selection spread across the genome has been shown in *Chapter 3* (Wong *et al.*, 2020).

4.3 Materials and Methods

4.3.1 Integrating the linkage map with genomic sequences

To generate an integrated map, markers used in linkage mapping in Chapman *et al.* (2013) were blasted against the *S. squalidus* reference genome. After identifying the locations of all the markers on the genomic scaffolds (ones with exact matches and no duplicates) and removing all other ambiguous matches, the linkage map positions of the markers (centimorgan, cM) were plotted against their respective genomic position on each of the scaffolds (base, b). Recombination rate was then calculated by dividing the difference in linkage map positions by the difference in genomic scaffold positions between each pair of markers closest to one another. The resulting recombination rates were plotted against positions on genomic scaffolds to look at the distribution of recombination rates along each genomic scaffold. Recombination rates over 40 cM/Mb were treated as errors and discarded.

4.3.2 Analyses of polymorphisms in nextRAD markers and differentially expressed genes

Two polymorphism summary statistics, F_{ST} (Weir & Cockerham, 1984) between *S. aethnensis* and *S. chrysanthemifolius* and nucleotide diversity (π) for each of the species, were calculated for a set of markers for comparisons with the integrated map and recombination rates. Markers from different sources were used: (a) 1769 nextRAD markers which have less than 50% missing data, from *Chapter 3* (Wong *et al.*, 2020); (b) 90 differentially expressed genes used in linkage mapping in Chapman *et al.* (2013) and generation of the integrated map in this study; and (c) differentially expressed genes with fold changes less than -2 or more than 2, and an adjusted P-value of less than 0.01, between the top and bottom transplant gardens in Walter *et al.* (2020b). F_{ST} and π for (a) and (b) were obtained from the respective previous studies. For markers in (c), F_{ST} between the two species and π in each species were estimated for all sites using ProSeq3 (Filatov, 2009), using existing transcriptomic sequences of each of the target differentially expressed genes from 16 *S. aethnensis* and 14 *S. chrysanthemifolius* individuals from Chapman *et al.* (2016), and genotypes used in the

transplant garden ('Sa' and 'Sc' samples) in Walter *et al.* (2020b) (Table 4.2). F_{ST} and π of all three types of markers were plotted along their positions on the genomic scaffolds in the same graph.

Table 4.2. List of samples used in estimating polymorphisms in differentially expressed genes from Walter *et al.* (2020b) (marker group c); and genome scans for selective sweeps. Sample names starting with 'Sa' or 'Sc' are genotypes of plants used in the transplant experiment in Walter *et al.* (2020b), hence no elevation data is shown. Other samples in 'Marker polymorphism' were existing transcriptomic data from Chapman *et al.* (2013); while sequences in 'Genome scan' were obtained from Filatov (unpublished).

Species	Sample name	Elevation (m)	Marker polymorphism	Genome scan
S. aethnensis	Pro1-4b	1810	✓	
	Pro1-34	1810	✓	
	Pro2-1	2060	✓	✓
	Pro2-2	2060	✓	✓
	Pro2-10	2060		✓
	Pro2-21	2060	✓	✓
	Pro2-22a	2060		✓
	Pro2-25	2060	✓	✓
	Pro2-20	2060	✓	
	Pro2-9	2060	✓	
	Et1-10	2105	✓	
	Et1-16	2105	✓	
	Et1-22	2105	✓	
	Et1-25A	2105	✓	
	Et3-2	2285	✓	
	Sa35	/	✓	
	Sa63	/	✓	
	Sa69	/	✓	
S. chrysanthemifolius	Nic1-21	230	✓	
	Ran1-5a	755	✓	✓
	Ran1-37a	755	✓	✓
	Ran1-43a	755	✓	✓
	Ran1-5a	755	✓	
	Rans-38	755	✓	
	Nic2-17	850	✓	✓
	Nic2-23	850	✓	✓
	Bro1-3	850	✓	
	Lin1-14a	925	✓	
	Lin1-17	925	✓	
	Nic2-34	950	✓	✓
	Sc11	/	✓	
	Sc18	/	✓	

In each π graph (one per linkage group), linear and quadratic regression were fitted to identify the better fitting one using the R package *stats* (R Core Team, 2013). The better-fitting regression type was determined by a higher correlation coefficient (R^2), and used to test whether π significantly decreased in regions with low recombination. Besides, recombination rates were plotted against π for each linkage group and all linkage groups combined to test their correlation.

4.3.3 Filtering of raw reads, SNP calling and sliding window analyses in the *S. aethnensis* and *S. chrysanthemifolius* genome sequences

In order to better understand polymorphism patterns and look for signals of selective sweeps in the *Senecio* genome, raw genomic sequences of six *S. aethnensis* and six *S. chrysanthemifolius* were obtained from previous genome sequencing projects (Table 4.2; Filatov, unpublished).

Low quality reads were first removed using default options in *trimmomatic* v0.36 (Bolger, Lohse, & Usadel, 2014). Filtered reads were then aligned to the indexed reference *S. squalidus* genome using *BWA mem* v0.7.15 (Li & Durbin, 2010). *SAMtools* v0.1.19 (Li *et al.*, 2009) was used to create and sort *.bam* fies. Duplicates were marked and removed using *Picard* v2.0.1 (Broad Institute, 2019). BCF files containing genotype likelihoods were obtained using the *mpileup* function in *SAMtools*. Using *BCFtools* within *SAMtools*, SNPs were then called using the *call* function, filtered using the *vcfutils.pl varFilter* function and finally exported as a VCF file. The genomic scaffold corresponding to the linkage groups identified by Chapman *et al.* (2016) were extracted into a new VCF file, with indels removed using *VCFtools* v0.1.13 (Danecek *et al.*, 2011).

Non-overlapping sliding window analyses for F_{ST} between the two species, π and Tajima's D for each species were carried out using *VCFtools.* Overlapping window analyses for F_{ST} and π were also carried out using the same tool. Window size for non-overlapping window analyses was 10,000 b; whereas the window size and step size for overlapping window analyses were 10,000 b and 2,000 b respectively.

4.3.4 Genomic scans for signals of selective sweeps and subsequent analyses of selection strength and age

Four software tools were used to do genomic scans to look for signals of selective sweeps in each of the species. In *RAiSD* (Alachiotis & Pavlidis, 2018), sliding window size (-w; in terms of number of SNPs and not bases on the genomic scaffolds) was set to 100 and the probability value (-A) was set to 0.995. In *OmegaPlus* (Alachiotis *et al.*, 2012), minimum window size (-minwin) and maximum window size (-maxwin) for computing linkage disequilibrium were set to 100b and 1,000b respectively. Grid size (-grid) was set to 8,000, meaning the scaffold would be divided into 8,000 equal fragments for analyses. This grid size resulted in fragments ranging from 6,541 to 9,895b in the scaffolds, comparable to the 10,000b windows used in *Chapter 3* (section 3.4.6). In *SweeD* (Pavlidis *et al.*, 2013), grid size (-grid) was set to 8,000 In *SweepFinder2* (DeGiorgio *et al.*, 2016), grid size (-s) was also set to 8,000, so that the results of *SweeD*, *OmegaPlus* and *Sweepfinder2* were directly comparable. Regions that have selection scores in the top 0.5% in *OmegaPlus*, *SweeD*, *SF2*, and top 0.1% in RAiSD were treated as outliers.

Analysing the outliers, common outliers between *S. aethnensis* and *S. chrysanthemifolius* in each software tool except *RAiSD*, and common outliers among software tools in each species were identified for each genomic scaffold. *RAiSD* was not included in this analysis because its approach in determining sliding windows is based on number of SNPs instead of bases and the use of overlapping windows resulted in incomparable windows with the other three software tools, which all have the same grid size hence windows.

To compare the strength and age of selective sweeps in outliers of the genome scans, and performance of software tools, selection coefficients and age of sweeps were estimated as follows. Selection coefficient (s) was estimated by $s = r*\ln(2N_e)/\alpha$ (Durrett & Schweinsberg, 2004), where α was outputted from the two software tools that utilise the CLR approach (*SweeD* and *SweepFinder2*; Nielsen et al., 2005), r represents the recombination rate, N_e represents the effective population size. The age of selective sweep (T) was then estimated using $T = 2*\ln(2N_e)/s$ (Stephan, Wiehe, & Lenz, 1992), which corresponds to $T = 2\alpha/r$ after combining the above two formulas.

Strong peaks in the outliers in each software tool and common peaks among them in each species were also identified and compared with their corresponding polymorphism patterns generated in section *4.3.3.* The three groups of markers, used in section *4.3.2* above, which are located in the regions with strong signals of selection were also identified and their gene annotation (GO) terms were obtained from *Uniprot* (https://www.uniprot.org/) to pinpoint potential genes that were under selection. As there were likely false-positive sweep signals, polymorphism patterns of the selected groups of sweep signals were compared and those with a high confidence of being true sweeps were identified using the following criteria. The groups were considered true sweeps if: (a) π was low or show a V-shape valley; (b) F_{ST} was moderate to high for cases in which there was sweep signal in only one species; (c) F_{ST} was low or high for cases in which there were sweep signals in both species.

4.3.5 Gene Ontology enrichment analysis

To test if any GO terms are over-represented, GO enrichment analyses were carried out using the online tool from The Gene Ontology Resources (www.geneontology.org), with genes identified in the previous section. This GO enrichment analysis tests if a certain GO term is over-represented compared to all the GO terms of a reference species. As neither *S. aethnensis* nor *S. chrysanthemifolius* has all their genes annotated, three model species were chosen instead: *Helianthus annuus* (common sunflower)*, Zea mays* (maize) and *Gossympium hirsutum* (upland cotton). These species were chosen as it is likely that different species have different sets of genes (hence difference set of GO terms) which are different to Etnean *Senecio*'s to different extents. The analyses were carried out with Fisher's Exact Test and calculation of false discovery rate (FDR).

4.4 Results

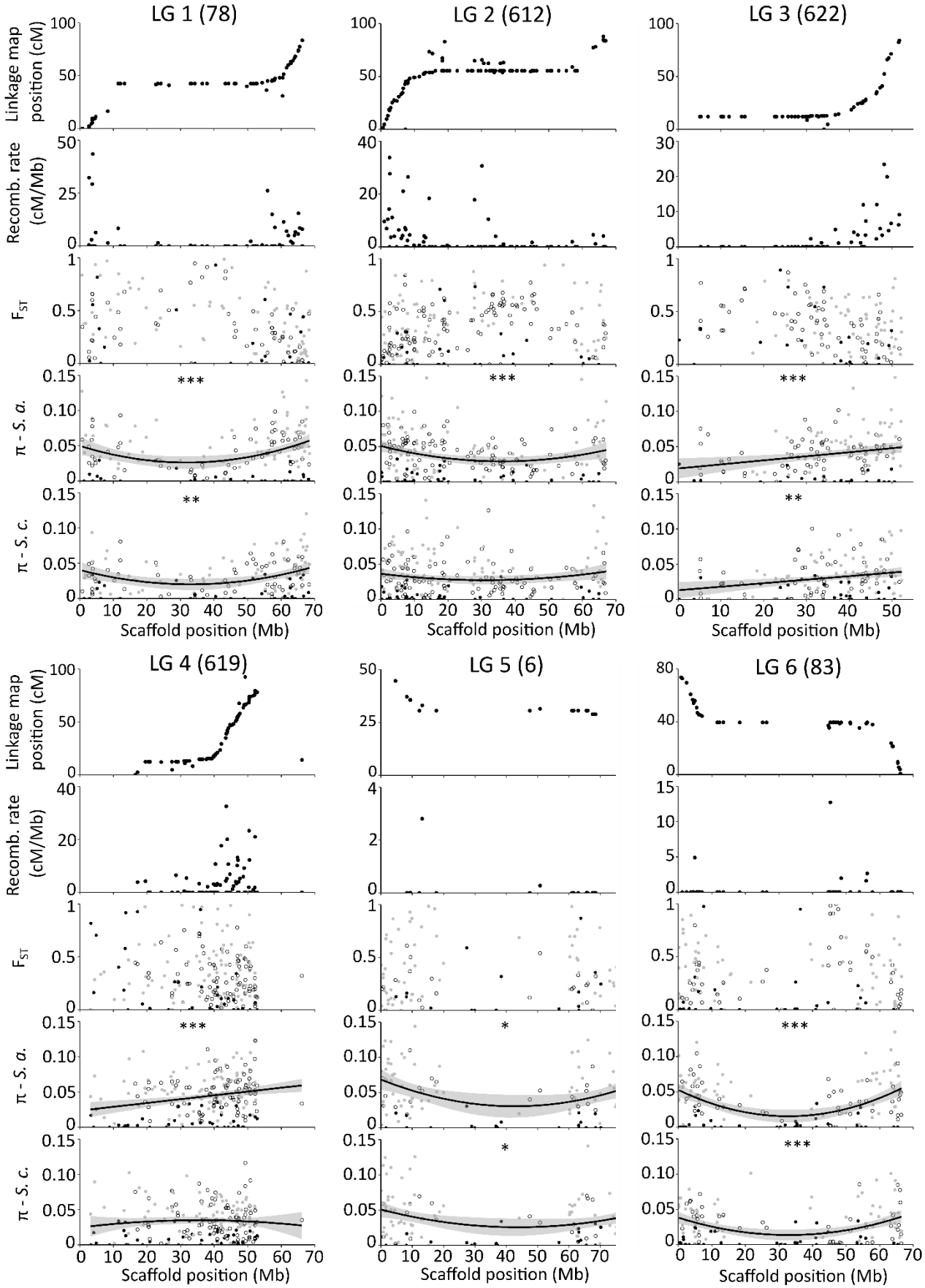
After calling and filtering SNPs and removing all indels, a total of 22,461,434 SNPs in ten genomic scaffolds were left for analyses. The tenlinkage groups in Chapman *et al.* (2016) matched to these ten genomic scaffolds of the *S. squalidus* genome, with number of SNPs ranging from 1,742,806 to 2,398,778 SNPs per scaffold and SNP density ranging from 0.033 to 0.038 SNPs/b (Figure 4.1).

Table 4.1. Length of scaffolds, number of SNPs and SNP density in each scaffold studied.

LG	Scaffold	Length of scaffold (b)	No. of SNPs	SNP density (/b)
1	78	68,463,240	2,398,778	0.035
2	612	67,696,633	2,484,653	0.037
3	622	52,330,192	1,757,164	0.034
4	619	52,882,715	1,742,806	0.033
5	6	79,160,621	3,000,805	0.038
6	83	66,670,782	2,355,818	0.035
7	620	67,673,870	2,329,662	0.034
8	621	57,835,927	1,910,353	0.033
9	538	59,368,209	2,282,327	0.038
10	462	59,330,977	2,199,068	0.037
All	All	631,413,166	22,461,434	0.036

4.4.1 Correlation between recombination rate and DNA polymorphism

Since no chromosome-level genomic data is available, it is assumed that the centromeres of chromosomes lie somewhere in the low-recombining region (flat part) in each linkage group of the integrated map (Figure 4.1). In general, F_{ST} between *S. aethnensis* and *S. chrysanthemifolius* is high in frequently-recombining regions but there are no obvious peaks (Figure 4.1).


LG 1 (78)
LG 2 (612)
LG 3 (622)
LG 4 (619)
LG 5 (6)
LG 6 (83)
Linkage map position (cM)
Recomb. rate (cM/Mb)
F_{ST}
π - S. a.
π - S. c.
***
**
*
Scaffold position (Mb)

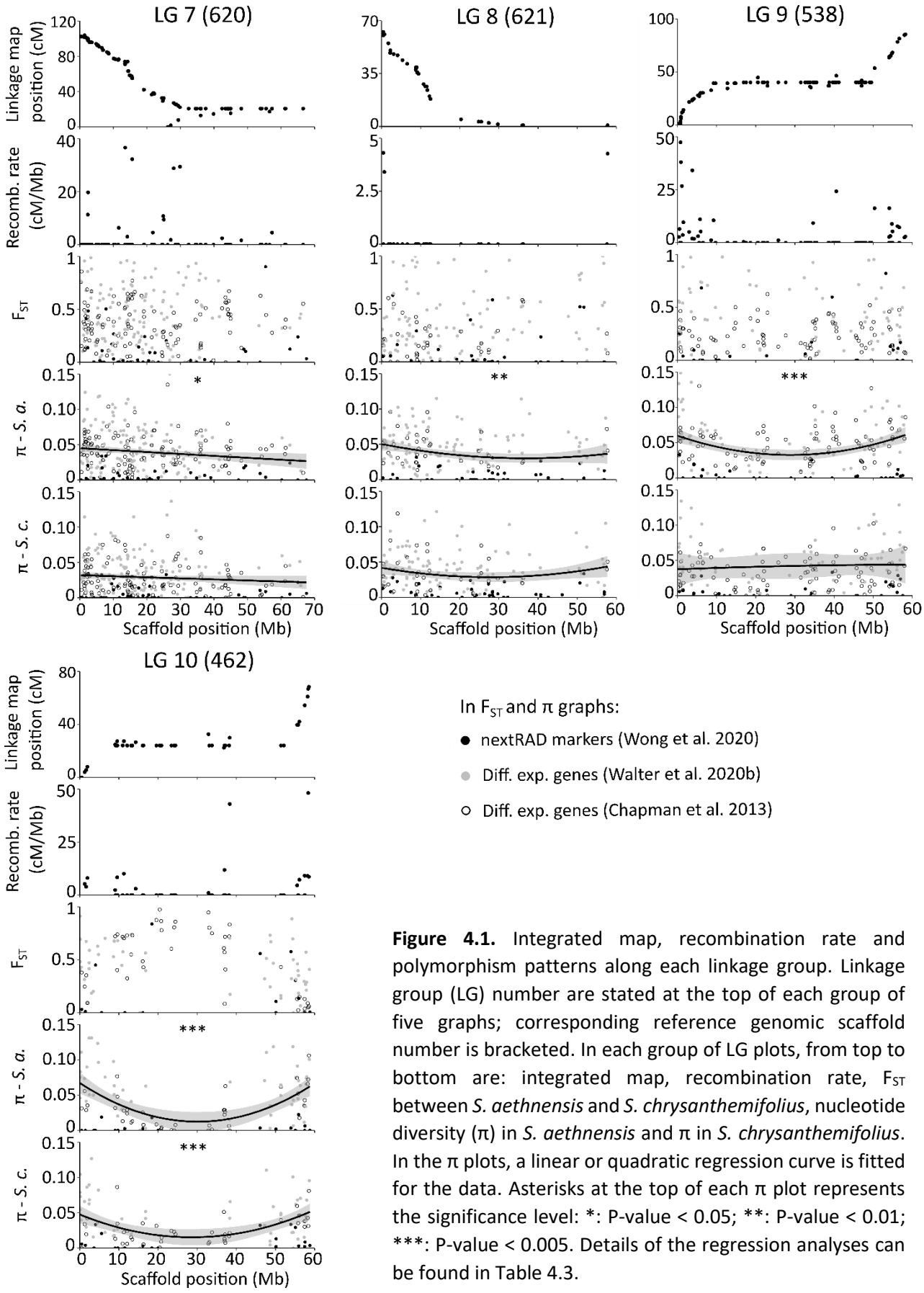


Figure 4.1. Integrated map, recombination rate and polymorphism patterns along each linkage group. Linkage group (LG) number are stated at the top of each group of five graphs; corresponding reference genomic scaffold number is bracketed. In each group of LG plots, from top to bottom are: integrated map, recombination rate, F_{ST} between *S. aethnensis* and *S. chrysanthemifolius*, nucleotide diversity (π) in *S. aethnensis* and π in *S. chrysanthemifolius*. In the π plots, a linear or quadratic regression curve is fitted for the data. Asterisks at the top of each π plot represents the significance level: *: P-value < 0.05; **: P-value < 0.01; ***: P-value < 0.005. Details of the regression analyses can be found in Table 4.3.

Regression analyses for nucleotide diversity (π) to test whether π was lower in low-recombining regions are shown in Table 4.3 and Figure 4.1. Quadratic regression explained more of the variances in LG 1, 2, 5, 6, 8, 9 and 10, where there were high-recombining regions at both ends of the linkage groups; whereas linear regression explained more of the variances in LG 3 and 7, where there was only highly-recombining region on one side of their integrated maps. In LG 4, linear regression fitted the *S. aethnensis* data better but quadratic regression fitted the *S. chrysanthemifolius* data better. Regression analysis results showed that π was significantly lower in low-recombining regions (positive correlation) in all linkage groups in *S. aethnensis*, but this positive correlation was only significant in LG 1, 3, 5, 6 and 10 in *S. chrysanthemifolius*.

Correlation between local recombination rate and π was further tested (Table 4.4, Figure 4.2). The two species were not significantly correlated overall among all linkage groups in both species. In *S. aethnensis,* recombination rate and π were significantly and positively correlated in LG 1, 2, 3, and 10; while in *S. chrysanthemifolius*, the two were only significantly and positively correlated in LG 3 and 10.

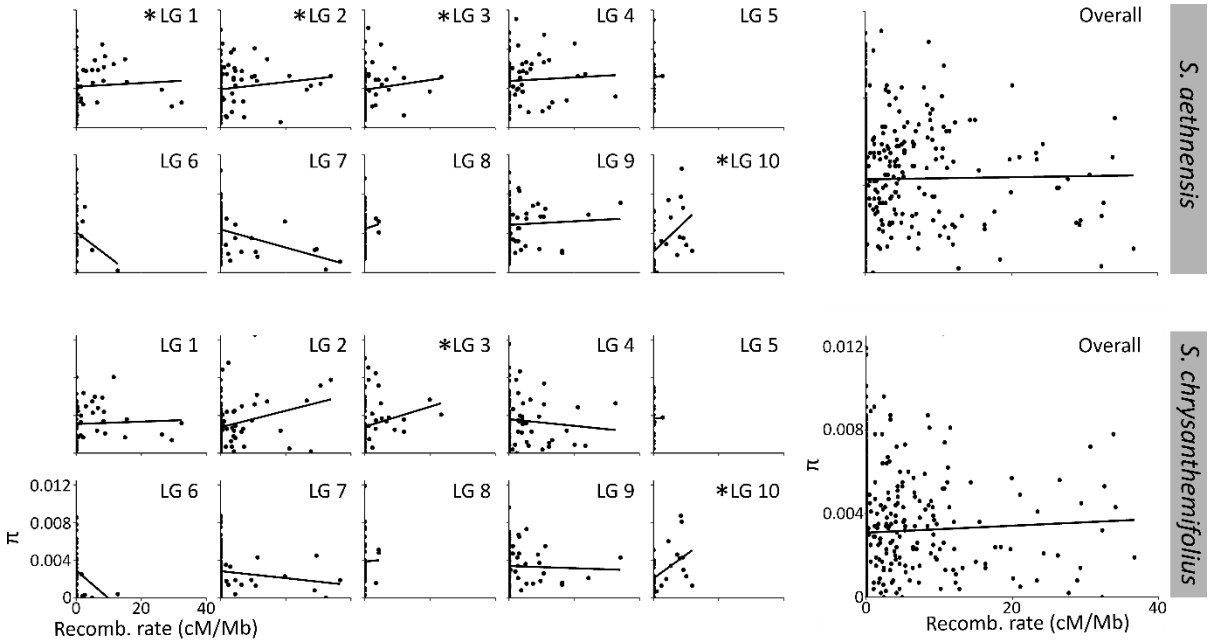


Figure 4.2. Relationships between nucleotide diversity (π) and recombination rate (cM/Mb) in each linkage group and overall among all linkage groups in *S. aethnensis* and π in *S. chrysanthemifolius*. Asterisks (*) next to the linkage group numbers indicates a significant correlation between the two indices. Statistics of the correlation analysis can be found in Table 4.4.

Table 4.3. Results of the regression analyses on nucleotide diversity (π) in each species. Better-fitting models were selected based on adjusted R^2 and are underlined and in *italic*.

LG	*S. aethnensis*								*S. chrysanthemifolius*							
	Linear model				Quadratic model				Linear model				Quadratic model			
	Multiple R^2	Adjusted R^2	F-stat	P-value	Multiple R^2	Adjusted R^2	F-stat	P-value	Multiple R^2	Adjusted R^2	F-stat	P-value	Multiple R^2	Adjusted R^2	F-stat	P-value
1	0.0152	0.0090	2.4670	0.1182	*0.0894*	*0.0779*	*7.8040*	*0.0006*	0.0055	-0.0007	0.8833	0.3487	*0.0633*	*0.0516*	*5.3760*	*0.0055*
2	0.0093	0.0049	2.1300	0.1458	*0.0487*	*0.0403*	*5.8120*	*0.0035*	0.0001	-0.0043	0.0295	0.8637	*0.0144*	*0.0057*	*1.6570*	*0.1930*
3	*0.0563*	*0.0499*	*8.8800*	*0.0034*	0.0594	0.0467	4.6720	0.0108	*0.0634*	*0.0571*	*10.0800*	*0.0018*	0.0634	0.0507	5.0080	0.0079
4	*0.0589*	*0.0545*	*13.4500*	*0.0003*	0.0592	0.0505	6.7380	0.0015	0.0011	-0.0035	0.2449	0.6212	*0.0049*	*-0.0044*	*0.5278*	*0.5907*
5	0.0093	0.0019	1.2650	0.2626	*0.0612*	*0.0472*	*4.3700*	*0.0145*	0.0116	0.0043	1.5880	0.2098	*0.0480*	*0.0338*	*3.3810*	*0.0369*
6	0.0002	-0.0064	0.0347	0.8524	*0.1571*	*0.1458*	*13.8900*	*0.0000*	0.0002	-0.0064	0.0329	0.8562	*0.0993*	*0.0872*	*8.2100*	*0.0004*
7	*0.0207*	*0.0174*	*6.1600*	*0.0136*	0.0224	0.0156	3.3190	0.0376	*0.0084*	*0.0049*	*2.4510*	*0.1185*	0.0112	0.0044	1.6410	0.1956
8	0.0315	0.0260	5.7560	0.0175	*0.0526*	*0.0418*	*4.8870*	*0.0086*	0.0004	-0.0052	0.0737	0.7863	*0.0246*	*0.0136*	*2.2230*	*0.1113*
9	0.0006	-0.0047	0.1107	0.7397	*0.0864*	*0.0767*	*8.8460*	*0.0002*	0.0010	-0.0043	0.1847	0.6679	*0.0010*	*-0.0097*	*0.0953*	*0.9091*
10	0.0000	-0.0088	0.0006	0.9811	*0.1811*	*0.1666*	*12.4900*	*0.0000*	0.0062	-0.0025	0.7101	0.4012	*0.1214*	*0.1059*	*7.8080*	*0.0007*

Table 4.4. Results of correlation analysis between recombination rate and nucleotide diversity (π) for each linkage group in each species.

LG	*S. aethnensis*				*S. chrysanthemifolius*			
	Multiple R^2	Adjusted R^2	F-stat	P-value	Multiple R^2	Adjusted R^2	F-stat	P-value
all	0.0007	-0.0009	0.4464	0.5043	0.0003	-0.0013	0.2073	0.6490
1	0.0719	0.0547	4.1810	0.0458	0.0168	-0.0014	0.9215	0.3414
2	0.0638	0.0541	6.5460	0.0121	0.0006	-0.0098	0.0599	0.8071
3	0.0771	0.0609	4.7640	0.0332	0.1734	0.1589	11.9600	0.0010
4	0.0147	0.0019	1.1480	0.2873	0.0018	-0.0111	0.1418	0.7075
5	0.0091	-0.0460	0.1650	0.6894	0.0429	-0.0102	0.8075	0.3807
6	0.0843	0.0635	4.0490	0.0503	0.0283	0.0062	1.2820	0.2637
7	0.0215	0.0120	2.2790	0.1342	0.0236	0.0142	2.5150	0.1158
8	0.0220	-0.0096	0.6961	0.4105	0.0015	-0.0308	0.0453	0.8329
9	0.0214	0.0097	1.8190	0.1811	0.0000	-0.0121	0.0000	0.9972
10	0.1219	0.0975	4.9980	0.0317	0.1039	0.0790	4.1720	0.0485

Table 4.5. Threshold values of outliers in each of the species in each software tools that detect selective sweeps. A = *S. aethnensis*; C = *S. chrysanthemifolius.*

		OmegaPlus (ω)		SweeD (Likelihood)		SweepFinder2 (LR)		RAiSD (μ)	
LG	Scaffold	A	C	A	C	A	C	A	C
1	78	26.22	30.12	17.90	40.72	163.26	221.42	15.38	25.70
2	612	19.52	27.00	7.85	43.78	533.62	224.27	19.21	22.20
3	622	18.27	33.38	7.17	35.61	177.37	451.28	17.29	26.41
4	619	15.79	28.29	18.31	14.84	178.32	199.62	19.40	19.44
5	6	19.04	35.66	12.46	15.55	262.56	201.90	20.20	19.47
6	83	19.11	20.95	7.17	21.04	358.35	473.84	19.22	22.81
7	620	19.38	31.14	7.53	32.76	277.61	286.94	17.40	25.72
8	621	13.08	21.47	21.59	35.27	243.09	222.81	17.98	31.08
9	538	26.33	23.50	6.47	19.04	180.37	187.86	16.68	20.30
10	462	21.16	27.72	29.58	17.47	118.94	347.28	16.83	20.06

Table 4.6. Number of common outliers of selective sweep signals between *S. aethnensis* and *S. chrysanthemifolius* in each software tool, and those between software tools in each species. A total of 40 outliers are present in each linkage group. There is no common outliers among all three software tools.

		Between *S. aeth.* and *S. chry.*			*S. aethnensis*			S. chrysanthemifolius		
LG	Scaffold	Omega+	SweeD	SF2	Between Omega+/SweeD	Between Omega+/SF2	Between SweeD/SF2	Between Omega+/SweeD	Between Omega+/SF2	Between SweeD/SF2
1	78	3	0	1	0	0	1	1	0	3
2	612	6	0	0	0	0	1	0	0	3
3	622	1	1	0	0	0	1	1	0	0
4	619	0	1	1	0	0	2	0	0	2
5	6	2	0	5	2	0	0	1	0	0
6	83	3	0	0	0	0	0	0	0	0
7	620	3	3	5	0	1	0	0	0	0
8	621	2	1	1	0	1	0	0	0	3
9	538	1	1	0	0	0	0	0	0	1
10	462	3	0	0	1	0	9	0	0	0

4.4.2 Sliding windows of polymorphisms and selective sweep scans in genome sequences

Distributions of DNA polymorphism summary statistics (Tajima's D and π for each species, and F_{ST} between the two species) are shown in Figure 4.3. In general, range of values of Tajima's D was wider in *S. chrysanthemifolius* than that in *S. aethnensis*, while π was similar between the two species and F_{ST} was variable. There were 40 outliers of selective sweep signals per linkage group in *OmegaPlus*, *SweeD* and *SweepFinder2* (*SF2*) (due to grid size of 8,000), and between 975 and 1729 outliers per linkage group in *RAiSD*. The signals of selective sweeps were very widespread within and across all linkage groups. *OmegaPlus* did not produce any obvious peaks in signals and only had scattered outliers across different linkage groups. In *SweeD* and *SF2*, in which they both estimate sweep signals based on SFS, occasionally shared strong sweep signals. *RAiSD* revealed many peaks in sweep signals (comprised of many outliers in each peak) and shared many outlier peaks with the other three software tools, as it integrated all three signatures of sweeps which are used by them.

Although there were sweep signals spread over each linkage group, they differed in the abundance of signals (Figure 4.4) as well as the threshold values of sweep signal outliers in the software tools (Table 4.5). Most linkage groups had widespread sweep signals, whereas some had more localised signals, such as LG 3 and 9. Sweep signals were generally stronger in *S. chrysanthemifolius* than in *S. aethnensis* – threshold values were higher in *S. chrysanthemifolius* except for LG 9 in *OmegaPlus*; LG 4 and 10 in *SweeD*; LG 2, 5, 8 in SF2; and LG 5 in *RAiSD*. Among the 40 outliers in *OmegaPlus*, *SweeD* and *SweepFinder2*, there were few common outliers between the two species (six or fewer; Table 4.6). In each species, there were few common outliers between each pair of software tools (nine or fewer) and none among the three (Table 4.7). Adjacent windows sometimes had outliers in different software tools.

Sweep signals also corresponded to various different patterns in polymorphisms (Table 4.8, Figure 4.3). In the selected regions with sweep signals, only some of them were deemed to have a high confidence of being true sweeps according to the criteria set out in section *4.3.4*. These regions are highlighted in Table 4.8. Out of the 57 regions analysed, 32 were deemed likely as true sweeps. The remaining regions likely had other processes that led to sweep-like patterns to be detected by each software tools or the resolution of data was not high enough to make definitive deductions.

4.4.3 Strength and age of chosen selective sweeps

LG 4 (Scaffold 619) was chosen to evaluate the performance of *SweeD* and *SF2* in analysing common sweeps, through comparing selection coefficients and ages of common sweeps yielded from their outputs. LG 4 was chosen because it resulted in equal numbers of common sweeps between the two software tools (two each) thus focus on software tools was not biased. LG 7 (Scaffold 620) was chosen to evaluate the strengths and ages of common sweeps (using α estimates from the two software tools; see section *4.3.4* for details) between *S. aethnensis* and *S. chrysanthemifolius*. LG 7 was chosen because it is the only linkage group that had more than one common sweep between species in both software tools. In this study, average recombination rate for LG 4 and LG 7 were 3.42 x 10^{-06} cM/b and 1.91 x 10^{-06} cM/b respectively. Effective population sizes of 18,531 and 12,770 were used for *S. aethnensis* and *S. chrysanthemifolius* respectively (Filatov *et al.*, 2016).

Selection coefficients (*s*) and age of sweeps (T) are summarised in Table 4.7. *s* and T estimates were quite different in common sweeps between *SweeD* and *SweepFinder2* in LG 4 (Scaffold 619): *s* ranged between 0.08 to 1.12 in *SweeD* and 0.39 to 1.76 in *SF2* (*s* being larger than one is likely due to overestimated recombination rate used in the estimations; discussed below); whereas T (generations) ranged from 151.90 to 253.52 in *SweeD* and 11.98 to 52.70 in *SF2* (Table 4.7).

For common sweeps between *S. aethnensis* and *S. chrysanthemifolius* in LG 7 (Scaffold 620), *SweeD* yielded similar estimates in *s* and T in the two species, with *s* ranging from 0.02 to 0.03 and T ranging from 608.44 to 1245.90 generations. In two out of three common sweeps in this scaffold, *SweeD* yielded the same α estimates for both species [hence the same *s* and T(generations)]. For common sweeps between the two species in the same scaffold, *SF2* yielded different *s* and T(generations) estimates for each species. In *S. aethnensis*, *s* ranged from 0.23 to 1.02 and T(generations) ranged from 20.67 to 91.17; whereas in *S. chrysanthemifolius*, *s* ranged from 0.34 to 1.08 and T(generations) ranged from 18.77 to 58.94.

4.4.4 Genes harboured in regions with strong putative selective sweeps

Various markers from the nextRAD dataset and differentially expressed genes from Chapman *et al.* (2013) and Walter *et al.* (2020b) matched to the regions with strong sweep signals in both species (Table 4.8). Regarding biological functions, many genes related to metal ion binding and transport were present in regions with strong sweep signals, including zinc (LG 2, 8, 9), iron and heme (LG 7, 8), and calcium (LG 7) ions for either or both species; while those involving magnesium (LG1, 7), and manganese (LG 3) ions were found only for *S. chrysanthemifolius*. Genes related to defence and stress were also found, including those related to lignin biosynthesis and cinnamyl alcohol (LG 2, 4), defence (LG 10), defence against fungus (LG 3, 7), wounding (LG 7), osmotic stress and water deprivation (LG, 3, 7), response to biotic stimuli (LG 4), and jasmonic acid and salicylic acid metabolism (LG 3). Several genes were related to growth and development, including shoot and root development (LG 3, 10), cell proliferation (LG 10), cell fate (LG 4), abscisic acid and cuticle (LG 3), and rubisco, melatonin and photoperiodism (LG 4). Interestingly, the low-elevation *S. chrysanthemifolius* showed regions with strong signals for genes involving ozone and UV-B response (LG 7), and embryo development ending in seed dormancy (LG 3, 7, 10), further discussed in the following section. Last but not least, LG 4 was the only linkage group involving genes related to reproduction for both species, including those related to flavanol, aromatic compounds, flower development, anther and fruit dehiscence, and pollen development.

Regarding cellular processes, many cellular processes were involved, without clear links to adaptation. However, there were some processes with more than one gene involved. These include protein ubiquitination (ubiquitin carboxyl-terminal hydrolase in LG 1 and two E3 ubiquitin-protein ligase PUB 23-like in LG 4), and pectin process (pectinesterase 2-like in LG 3 and pectin lyase-like superfamily protein in LG 4).

4.4.5 Gene Ontology enrichment analyses

GO enrichment analysis revealed many over-represented GO terms with apparent functions related to adaptation to Mount Etna (see list of over-represented GO terms, their fold changes and P-values in Table 4.9 to Table 4.11). In terms of biological functions (Table 4.9), over-represented GO terms include those related to:

- aromatic compounds (*H. annuss*: GO:0006725; *Z. mays*: GO:0019439);
- taxis (*Z. mays*: GO:0042339, GO: 0050918, GO:0006935);
- pollen tube (*Z. mays*: GO:0010183);
- defence response (*Z. mays*: Go:0006952; *G. hirsutum*: GO:0098542, GO:0006952);
- response to biotic stimulus/ other organism (*G. hirsutum*: GO:0009607, GO:0043207, GO:0050896, GO:0051707);
- response to stress (*G. hirsutum*: GO:0006950);
- phosphorylation (*G. hirsutum:* GO:0016310); and
- sucrose (*G. hirsutum*: GO:0005985, GO:0015770).

Over-represented GO terms regarding molecular functions (Table 4.10) include:

- ion binding (*Z. mays*/*G. hirsutum*: GO:0043167);
- iron ion binding (*G. hirsutum*: GO:0005506);
- heme ion binding (*G. hirsutum*: GO:0020037);
- metal ion binding (*G. hirsutum*: GO:0046914); and
- sucrose (*G. hirsutum*: GO:0008506, GO:0008515, GO:0009669).

Table 4.7. Comparisons of selection coefficient (*s*) and age of selected alleles (T) in selected outliers between the two software tools that utilise the CLR approach in each species, and those between the two species in each software tool. CLR and α are outputs from *SweeD* and *SweepFinder2*; r = average recombination rate for the corresponding LG; N_e is based on estimates from Filatov *et al.* (2016).

		Position	***SweeD***						***SweepFinder2***					
			CLR	**α (x10^{-4})**	**r/α**	***s***	**T (gen)**	**T (4N$_e$)**	**CLR**	**α (x10^{-4})**	**r/α**	***s***	**T (gen)**	**T (4N$_e$)**
Common sweeps between *SweeD* and *SF2* in LG4 (Scaffold 619)	In *S. aethnensis*	36491794	21.76	2.60	0.0132	0.14	151.90	0.0020	509.76	0.20	0.1670	1.76	11.98	0.0002
		36498405	72.99	4.33	0.0079	0.08	253.52	0.0034	461.07	0.22	0.1571	1.65	12.73	0.0002
	In *S. chrysanthemifolius*	47022845	14.84	2.80	0.0122	0.12	163.65	0.0032	310.72	0.61	0.0561	0.57	35.65	0.0007
		47029456	16.39	3.42	0.0100	0.10	200.03	0.0039	457.78	0.90	0.0380	0.39	52.70	0.0010
		Position	***S. aethnensis***						***S. chrysanthemifolius***					
			CLR	**α (x10^{-4})**	**r/α**	***s***	**T (gen)**	**T (4N$_e$)**	**CLR**	**α (x10^{-4})**	**r/α**	***s***	**T (gen)**	**T (4N$_e$)**
Common sweeps between *S. aeth.* and *chry*. in LG7 (Scaffold 620)	In *SweeD*	39383197	44.57	9.88	0.0019	0.02	1032.16	0.0139	50.90	9.88	0.0019	0.02	1032.16	0.0202
		39391657	39.62	5.82	0.0033	0.03	608.44	0.0082	49.97	5.82	0.0033	0.03	608.44	0.0119
		39408577	47.66	11.92	0.0016	0.02	1245.90	0.0168	53.36	9.63	0.0020	0.02	1006.64	0.0197
	In *SweepFinder2*	2683179	279.59	0.87	0.0219	0.23	91.17	0.0012	325.68	0.56	0.0339	0.34	58.94	0.0012
		5931866	583.62	0.22	0.0851	0.90	23.50	0.0003	580.72	0.18	0.1065	1.08	18.77	0.0004
		5940326	579.76	0.20	0.0965	1.01	20.73	0.0003	355.52	0.21	0.0910	0.92	21.99	0.0004
		6346412	610.07	0.20	0.0968	1.02	20.67	0.0003	352.51	0.42	0.0454	0.46	44.06	0.0009
		38943269	430.06	0.36	0.0532	0.56	37.58	0.0005	295.49	0.55	0.0348	0.35	57.52	0.0011

Table 4.8. Details of chosen groups of sweep signals in various software tools that are highlighted in Figure 4.3. Markers from the datasets nextRAD, Chapman *et al.* (2013), and Walter *et al.* (2020b) that lie in each region are listed (markers without matched genes are omitted). Regions that have high confidence of the sweep signals being true sweeps are highlighted in bold and underlined in the first five columns. In the column *Species*, A = *S. aethnensis*; C = *S. chrysanthemifolius.* In the column *Selected G.O. terms*, F = molecular function; P = biological process. G.O. terms that are potentially related to adaptation to environments are underlined.

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG1 (Scaffold78)									
0-5	RAiSD	C	V-shape	V-shape	1782996	OG0000905_2297	Nonaspanin	PWA74422.1	N/A
25-30	**SweeD**	**A,C**	**V-shape**	**low**	N/A	N/A	N/A	N/A	N/A
29-35	**SF2**	**A**	**V-shape**	**high**	29996628	OG0001992_4356	Oligopeptide transporter 5-like	XM_025133542.1	N/A
35-48	RAiSD	C	variable	high	N/A	N/A	N/A	N/A	N/A
55-58	OmegaPlus, RAiSD	A	V-shape	low	N/A	N/A	N/A	N/A	N/A
64-66	**SF2, RAiSD**	**C**	**V-shape**	**moderate**	64406626	OG0000919_2323	AAA-ATPase At5g57480-like	XM_023891043.1	F: ATP binding
					64435602	OG0006367_9056	Zinc finger protein 2-like	XM_022139363.1	N/A
					64646500	OG0005367_8056	At4g30420-like	XM_023880024.1	F: transmembrane transporter activity
					65018363	Hybrid_contig_6444	Pyruvate kinase, cytosolic isozyme-like	XM_022124904.1	F: magnesium/ potassium ion binding, pyruvate kinase activity; P: glycolytic process, phosphorylation
					65044495	Hybrid_contig_15982	OTU domain-containing protein 6B	XM_022141125.1	N/A

65064919	OG0017055_19743	Serine carboxypeptidase 24-like	XM_022141121.1	F: serine-type carboxypeptidase activity, carboxypeptidase activity; P: proteolysis
65104063	Hybrid_contig_1722	Ubiquitin carboxyl-terminal hydrolase 23	XM_025123198.1	F: peptidase activity, metal ion binding, cysteine-type peptidase activity; P: protein deubiquitination, proteolysis

*In 25-30Mb, sweep peak in A corresponds to a V-shaped valley in π in C; vice versa.

*The cluster of sweep peaks at 35-48Mb have similar heights as the group at ~10-20Mb but F_{ST} of the latter is conspicuously lower.

*In RAiSD, C had more prominent and more number of peaks than A.

*Highs and lows in μ_{chry} and F_{ST} patterns seem to correspond with each other.

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG2 (Scaffold612)									
10-15	**SF2, RAiSD**	**A,C**	**low + very high peak**	**high**	12117909	OG0001746_3891	Polygalacturonase	XM_025138694.1	N/A
25-30	SweeD	C	peak	high	N/A	N/A	N/A	N/A	N/A
30-32	SF2, RAiSD	A	moderate	V-shape	30202491	OG0002239_4815	Cinnamyl alcohol dehydrogenase 1-like	XM_023912785.1	F: zinc ion binding, cinnamyl-alcohol dehydrogenase activity, metal ion binding; P: oxidation-reduction process, lignin biosynthetic process
52-55	RAiSD	C	high	moderate	N/A	N/A	N/A	N/A	N/A
60-62	SF2, RAiSD	C	moderate	high	N/A	N/A	N/A	N/A	N/A
* Tajima's D is low in both species at around 30-50Mb, where there are no strong peaks in selection scans.									

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG3 (Scaffold622)									
4-7	OmegaPlus, SweeD	C	V-shape in A	peak	N/A	N/A	N/A	N/A	N/A
9-11	**RAiSD**	**A,C**	**V-shape**	**low**	N/A	N/A	N/A	N/A	N/A
20-22	RAiSD	A,C	peak	V-shape	N/A	N/A	N/A	N/A	N/A
33-39	**SweeD, SF2, RAiSD**	**C**	**variable**	**high**	34034201	Hybrid_contig_7533	mRNA-decapping enzyme subunit 2	XM_022157908.1	F: manganese ion binding, hydrolase activity; P: modulation by virus of host process, primary shoot apical meristem specification
					34089167	Hybrid_contig_2979	Guanylate kinase 2	XM_022121244.1	F: guanylate kinase activity, transferase activity; P: phosphorylation, regulation of developmental growth
					34089462	10645	Guanylate kinase 2	XM_011073219.2	see above
					34220381	Hybrid_contig_7180	Inactive ATP-dependent zinc metalloprotease FTSHI 4	XM_025121895.1	F: metalloendopeptidase activity; P: proteolysis, embryo development ending in seed dormancy, protein import into chloroplast stroma
					38045341	OG0006268_8957	Methylesterase 17-like	XM_022118412.1	P: salicylic acid/ jasmonic acid metabolic process
42-48	**OmegaPlus, SF2, RAiSD**	**A,C**	**V-shape**	**moderate**	43784953	Hybrid_contig_3003	Glycerol-3-phosphate acyltransferase	XM_025105451.1	F: transferase activity, transferring acyl groups; P: lipid metabolic process, phospholipid/ phosphatidylglycerol biosynthetic process
					43887341	Hybrid_contig_5047	Protein HLB1-like	XM_025141056.1	N/A

					43939591	OG0006884_9573	Aldehyde oxidase GLOX-like	XM_023886636.1	N/A
					44016679	OG0001640_3687	Lysophospholipase BODYGUARD 1	XM_022146331.1	P: response to osmotic stress, regulation of abscisic acid biosynthetic process, lateral root development, defense response to fungus, cuticle development, suberin biosynthetic process
					44453009	OG0004850_7539	Bidirectional sugar transporter SWEET4-like	XM_022171327.1	F: sugar transmembrane transporter activity; P: carbohydrate transport, carbohydrate transmembrane transport
					46284545	OG0012104_14793	LURP-one-related 4-like	XM_023897023.1	N/A
49-52	SF2, RAiSD	A,C	high	mdoerate	50881497	OG0003429_6118	Pectinesterase 2-like	XM_022174514.1	N/A

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG4 (Scaffold619)									
3-7	**SweeD, SF2, RAiSD**	**A,C**	**V-shape**	**mdoerate**	4414414	OG0019887_22575	Thaumatin-like protein 1	XM_023913496.1	F: hydrolase activity, acting on glycosyl bonds; P: defense response, response to biotic stimulus
12-21	**SweeD, SF2, RAiSD**	**A,C**	**V-shape**	**low**	13180325	OG0017646_20334	Flavone 3'-O-methyltransferase 1	XM_023901178.1	F: protein dimerization activity, transferase activity; P: lignin/ flavonol/ aromatic compound/ melatonin biosyntehtic process
					13746109	4514	Axial regulator YABBY 2	XP_024976757.1	F: metal ion binding, DNA binding; P: cell fate commitment, abaxial cell fate specification, multicellular organism development
					13883219	17126	Outer envelope protein 61	XM_022181763.1	P: protein transport, cytosol to endoplasmic reticulum transport
					14354108	8983	E3 ubiquitin-protein ligase PUB23-like	XM_019384623.1	F: ubiquitin-protein transferase activity; P: protein ubiquitination
					17137281	Hybrid_contig_3881	DEAD-box ATP-dependent RNA helicase 13	XM_025114817.1	F: RNA binding, RNA helicase activity, nucleic acid/ nucleotide/ ATP binding, hydrolase activity, hylicase activity; P: rRNA processing, ribosome biogenesis
					17631974	OG0000022_133	Pectin lyase-like superfamily protein	PWA89210.1	F: polygalacturonase activity, hydrolase activity; P: anther dehiscence, fruit dehiscence,

									pectin catabolic process, pollen sperm cell differentiation
					18854530	OG0015816_18504	ECERIFERUM 26-like	XM_022153510.1	F: transferase activity, transferring acyl groups other than amino-acyl groups; P: cell wall organization, pollen development, very long-chain fatty acid biosyntehtic process
					19434451	OG0000437_1353	Pathogen-related protein-like	XM_022179046.1	N/A
					19525749	Hybrid_contig_3946	Inactive leucine-rich repeat receptor-like protein kinase At3g28040	XM_022172780.1	F: protein kinase activity, ATP binding, protein binding; P: protein phosphorylation, nucleotide binding
					20415848	Hybrid_contig_6051	Transcription antitermination protein NusB	XM_025122166.1	N/A
					20467292	Hybrid_contig_1806	ruBisCO large subunit-binding protein subunit beta	XM_025119515.1	F: ATP binding; P: protein folding
					21537212	OG0000334_1085	E3 ubiquitin-protein ligase PUB23-like	XM_025120210.1	see above
					22351873	OG0014903_17592	Transcription repressor OFP15-like	XM_023907895.1	P: negative regulation of transcription, DNA-templated
22-28	**All**	**A,C**	**V-shape**	**high**	N/A	N/A	N/A	N/A	N/A
28-37	All	A,C	high	high	33559473	Hybrid_contig_2602	Ribonuclease E/G-like protein	XM_025112283.1	F: hydrolase activity, nuclease activity, ribonuclease activity, starch binding; P: chloroplast organization, RNA/ nucleic acid phosphodiester bond hydrolysis

					33877402	OG0005991_8680	Transcription initiation factor TFIID subunit 14b-like	XM_025130800.1	P: regulation of flower development, histone H4-K5 acetylation, regulation of timing of transition from vegetative to reproductive phase, regulation of photoperiodism, flowering
					34775206	OG0005244_7933	GbDFR mRNA for dihydroflavonol reductase	AB550240.1	N/A
					35326429	OG0011455_14144	Scopoletin glucosyltransferase-like	XM_019334280.1	F: UDP-glycosyltransferase activity, transferase activity, transferring glycosyl groups
37-39	OmegaPlus, SweeD, RAiSD	A,C	high	high	N/A	N/A	N/A	N/A	N/A
44-54	**All**	**A,C**	**variable**	**variable**	N/A	N/A	N/A	N/A	N/A
*The genes matched to the region 3-7Mb and 22-37Mb belong to sweep signals in A only.									
*In RAiSD, C had more prominent and more number of peaks than A.									

Scaffold position (Mb)	Software tool	Species involved	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG5 (Scaffold6)									
8-10	**SF2, RAiSD**	**C**	**V-shape**	**peak**	N/A	N/A	N/A	N/A	N/A
11-23	**All**	**A**	**variable**	**variable**	N/A	N/A	N/A	N/A	N/A
26-29	**SweeD, SF2, RAiSD**	**A,C**	**V-shape**	**peak**	N/A	N/A	N/A	N/A	N/A
37-45	All	A,C	high where SF2 peaks are	high where SF2 peaks are	N/A	N/A	N/A	N/A	N/A
69-80	All	A,C	moderate	variable	N/A	N/A	N/A	N/A	N/A
*In 69-80Mb, FST is high where there is a strong A peak in RAiSD, but lower in the rest of the cluster where there are mostly C peaks only.									
*In RAiSD, there are lots of selection signals - in some regions there are regions in which signals are either strong in A or C.									
LG6 (Scaffold83)									
6-8	All	C	high	moderate	N/A	N/A	N/A	N/A	N/A
20-22	**SF2, RAiSD**	**A,C**	**V-shape in A**	**V-shape**	N/A	N/A	N/A	N/A	N/A
40-43	**SweeD, SF2, RAiSD**	**A,C**	**V-shape**	**low**	N/A	N/A	N/A	N/A	N/A
45-47	**All**	**A,C**	**V-shape**	**high**	N/A	N/A	N/A	N/A	N/A
*At around 25Mb, there are no big selection signals, but π is very high in both species.									

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG7 (Scaffold620)									
5-8	SF2	A,C	moderate	low	N/A	N/A	N/A	N/A	N/A
16-19	RAiSD	A,C	moderate	V-shape	16916912	OG0019418_22106	Jasmone hydroxylase CYP71AT148	MG189934.1	F: iron ion binding, heme binding, oxidoreductase activity; P: oxidation-reduction process
					17248770	3756	Lysine-specific histone demethylase 1	XR_003068495.1	F: oxidoreductase activity, polyamine oxidase activity, DNA-binding; P: chromatin organization, polyamine catabolic process
					17321249	OG0000438_1355	Cysteine proteinase 15A-like	XM_022122003.1	F: cysteine-type peptidase activity, hydrolase activity; P: proteolysis
					17453593	OG0009997_12686	Phosphatidylinositol/ phosphatidylcholine transfer protein SFH6-like	XM_025113995.1	N/A
					17889237	OG0005856_8545	Calcium uniporter protein 4	XM_022118345.1	F: calcium channel activity, uniporter activity; P: mitochondrial calcium ion homeostasis, calcium ion transport
					18003310	OG0007138_9827	Cytochrome P450 CYP82D47-like	XM_022127991.1	F: iron ion binding, heme binding, monooxygenase activity; P: oxidation-reduction process
19-22	SF2, RAiSD	C	Small peak	moderate	18655598	OG0017108_19796	Inactive receptor kinase At5g67200	XM_023885300.1	F: protein kinase activity, ATP binding; P: protein phosphorylation

18694877	OG0013135_15824	Extensin-like	XM_025113940.1	N/A
18934846	OG0005422_8111	Peamaclein	XM_025137427.1	N/A
19059130	OG0007201_9890	UV-B-induced protein At3g17800	XM_025113754.1	P: response to ozone, response to wounding, response to UV-B
19074566	Hybrid_contig_2780	Adenine phosphori-bosyltransferase 1-like	XM_025114521.1	F: adenine phosphribosyltransferase, transferase activity; P: nucleoside metabolic process, adenine salvage
19138339	OG0001893_4169	Interferon-related developmental regulator 1-like	XM_022127792.1	N/A
19148422	Hybrid_contig_4610	Calcium-transporting ATPase 1-like	XM_025115314.1	F: magnesium transmembrane transporter activity phosphorylative mechanism; P: magnesium ion transport
19391046	OG0014604_17293	CCG-binding protein 1	XM_023656273.1	F: protein binding, mediator complex binding; P: embryo development ending in seed dormancy, defense response to fungus
19443613	OG0004354_7043	Ethylene-responsive transcription factor ERF019-like	XM_022115508.1	N/A
19570600	OG0014550_17239	GDSL esterase/lipase At1g33811	XM_023903550.1	F: hydrolase activity; P: lipid metabolic process, lipid catabolic process
19725575	OG0001402_3237	Sucrose transport protein SUC8-like	XM_022127128.1	F; sucrose, proton symporter activity; P: sucrose transport

					19771006	OG0010885_13574	ABC transporter G family member 25-like	XM_022137065.1	F: nucleotide binding, ATP binding, ATPase activity
					20876755	2542	Multidrug resistance protein 2-like	XM_022127210.1	N/A
31-34	**All**	**A,C**	**V-shape**	**V-shape**	N/A	N/A	N/A	N/A	N/A
36-38	All	A,C	high	V-shape	N/A	N/A	N/A	N/A	N/A
42-45	**SweeD, RAiSD**	**C**	**V-shape**	**moderate**	43405887	Hybrid_contig_7290	Pre-mRNA-splicing factor 38	XM_025113004.1	P: RNA splicing, mRNA processing
					43420805	Hybrid_contig_4439	Topless-related protein 3-like	XM_022118483.1	P: regulation of transcription, DNA-templated
					43507223	Hybrid_contig_8353	Protein phosphatase 2C 34	XM_022138140.1	F: hydrolase activity, metal ion binding; P: protein serine/threonine/ phosphoprotein phosphatase activity, response to water deprivation, regulation of microtubule binding, negative regulation of growth
					43743585	Hybrid_contig_8440	Protein FAM133	XM_023880550.1	N/A
53-65	**OmegaPlus, SweeD**	**A,C**	**variable**	**moderate**	N/A	N/A	N/A	N/A	N/A
65-68	**RAiSD**	**A,C**	**V-shape**	**moderate**	N/A	N/A	N/A	N/A	N/A

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG8 (Scaffold621)									
7-9	**SF2, RAiSD**	**C**	**V-shape**	**moderate**	N/A	N/A	N/A	N/A	N/A
13-15	**OmegaPlus, SF2, RAiSD**	**A,C**	**V-shape**	**low**	14069633	OG0011429_14118	Zinc finger protein CONSTANS-LIKE 16-like	XM_023904034.1	F: zinc ion binding; P: regulation of transcription, DNA-templated
					14091681	OG0000250_871	Geraniol 8-hydroxylase-like	XM_022180170.1	F: iron ion binding, heme binding, oxidoreductase activity, acting on paired donors, with incorporation of reduction of molecular oxygen, metal ion binding, monooxygenase activity; P: oxidation-reduction process
18-20	**All**	**A,C**	**peak in A**	**moderate**	N/A	N/A	N/A	N/A	N/A
22-28	**SF2, RAiSD**	**A,C**	**variable**	**low**	N/A	N/A	N/A	N/A	N/A
34-36	SweeD, SF2, RAiSD	A,C	peak	peak	N/A	N/A	N/A	N/A	N/A
42-44	RAiSD	A	moderate	low	N/A	N/A	N/A	N/A	
45-48	**SF2, RAiSD**	**C**	**V-shape**	**V-shape**	N/A	N/A	N/A	N/A	

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG9 (Scaffold538)									
8-11	**SF2, RAiSD**	**A,C**	**V-shape**	**V-shape**	906152	Hybrid_contig_7129	Pre-mRNA-splicing factor 18	XM_022185484.1	N/A
					935175	OG0009266_11955	40S ribosomal protein S24-1	XM_022179851.1	F: mRNA binding, structural constituent of ribosome; P: translation
14-16	SweeD, SF2	C	peak	V-shape	N/A	N/A	N/A	N/A	N/A
31-33	**All**	**A,C**	**V-shape**	**moderate**	N/A	N/A	N/A	N/A	N/A
33-35	SF2	A	V-shape	low	34299785	932	Zinc/iron permease	PWA39693.1	F: zinc/ metal ion transmembrane transporter activity; P: zinc ion transmembrane transport, metal ion transport
38-40	SweeD, SF2, RAiSD	C	high	high	N/A	N/A	N/A	N/A	N/A
55-58	**SweeD, SF2, RAiSD**	**A,C**	**V-shape**	**low**	N/A	N/A	N/A	N/A	N/A

Scaffold position (Mb)	Software tool	Species	π	F_{ST}	Markers in selection regions				
					Scaffold position (b)	Marker name	Gene	Gene accession no.	Selected G.O. terms
LG10 (Scaffold462)									
5-13	**SweeD, RAiSD**	**A,C**	**moderate + peaks**	**variable**	11490770	11609	CRAL-TRIO domain-containing protein C23B6.04c	XM_023872815.1	F: phospholipid transfer activity; P: intermembrane phospholipid transfer
40-42	SweeD, SF2, RAiSD	A	V-shape	V-shape	N/A	N/A	N/A	N/A	N/A
42-45	**RAiSD**	**C**	**wide V-shape**	**moderate**	N/A	N/A	N/A	N/A	N/A
53-56	**SweeD, SF2, RAiSD**	**C**	**V-shapes**	**variable**	53945587	15314	DNA replication licensing factor MCM2	XM_004299151.2	F: protein/ DNA/ ATP/ metal ion binding, helicase activity; P: DNA replication initiation, embryo development ending in seed dormancy, regulation of cell population proliferation, root development, viral process
					54193479	OG0000205_746	Xyloglucan glycosyltransferase 5	XM_025114308.1	F: transferase activity, xyloglucan 6-xylosyltransferase activity; P: xyloglucan biosynthetic/ metabolic process, root hair elongation
					54426844	OG0000772_2070	GDSL esterase/ lipase 1-like	XR_002855376.1	F: lipase activity, hydrolase activity, acting on ester bonds; P: lipid metabolic process

	54533392	OG0018186_20874	dnaJ homolog subfamily B member 7-like	XM_023878826.1	N/A
	55097899	7005	Protein phosphatase 2C 16-like	XM_022128825.1	F: cation binding, catalytic activity, protein serine/threonine phosphatase activity, metal ion binding; P; protein dephosphorylation
	55204110	OG0014864_17553	Thaumatin-like protein 1b	XM_009351412.2	P: defense response

*In 54-56Mb, π in A also showed V-shape and A showed some low peaks in the regions too.

Table 4.9. List of over-represented GO terms for biological processes using genes identified in Table 4.8 and annotations from three species (*Helianthus annus*, *Zea mays*, *Gossypium hirsutum*).

	Fold enrichment	Raw P-value	FDR P-value
Helianthus annus			
generation of catalytic spliceosome for second transesterification step (GO:0000350)	> 100	1.61E-07	4.68E-05
negative regulation of DNA duplex unwinding (GO:1905463)	> 100	4.11E-04	4.78E-02
regulation of DNA duplex unwinding (GO:1905462)	> 100	4.11E-04	4.66E-02
negative regulation of ATPase activity (GO:0032780)	> 100	4.11E-04	4.55E-02
negative regulation of helicase activity (GO:0051097)	> 100	4.11E-04	4.44E-02
negative regulation of DNA helicase activity (GO:1905775)	> 100	4.11E-04	4.34E-02
regulation of DNA helicase activity (GO:1905774)	> 100	4.11E-04	4.25E-02
spliceosomal conformational changes to generate catalytic conformation (GO:0000393)	> 100	4.51E-07	1.23E-04
nuclear retention of unspliced pre-mRNA at the site of transcription (GO:0071048)	> 100	1.47E-06	3.78E-04
nuclear mRNA surveillance of spliceosomal pre-mRNA splicing (GO:0071030)	> 100	1.47E-06	3.58E-04
intracellular mRNA localization (GO:0008298)	> 100	2.46E-06	5.72E-04
nuclear retention of pre-mRNA at the site of transcription (GO:0071033)	> 100	2.46E-06	5.45E-04
nuclear mRNA surveillance (GO:0071028)	> 100	6.08E-06	1.28E-03
nuclear RNA surveillance (GO:0071027)	> 100	6.08E-06	1.23E-03
RNA surveillance (GO:0071025)	> 100	6.53E-06	1.26E-03
mRNA splicing, via spliceosome (GO:0000398)	> 100	1.92E-13	8.92E-10
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	> 100	2.73E-13	6.34E-10
RNA splicing, via transesterification reactions (GO:0000375)	> 100	2.73E-13	4.23E-10
RNA splicing (GO:0008380)	> 100	7.95E-13	9.23E-10
RNA localization (GO:0006403)	> 100	1.22E-04	1.77E-02
nuclear-transcribed mRNA catabolic process (GO:0000956)	> 100	1.38E-04	1.95E-02
mRNA catabolic process (GO:0006402)	> 100	1.55E-04	2.06E-02
mRNA processing (GO:0006397)	> 100	2.33E-12	2.17E-09
RNA catabolic process (GO:0006401)	85.27	2.37E-04	2.98E-02
mRNA metabolic process (GO:0016071)	74.65	1.65E-11	1.27E-08
nucleobase-containing compound catabolic process (GO:0034655)	64.81	4.08E-04	4.85E-02
RNA processing (GO:0006396)	42.8	4.53E-10	3.00E-07
cellular protein-containing complex assembly (GO:0034622)	32.84	7.54E-05	1.25E-02
protein-containing complex assembly (GO:0065003)	30.08	9.77E-05	1.46E-02
protein-containing complex subunit organization (GO:0043933)	26.16	1.47E-04	2.01E-02
negative regulation of biological process (GO:0048519)	24.97	1.69E-04	2.18E-02
RNA metabolic process (GO:0016070)	24.02	1.42E-08	5.98E-06
cellular component assembly (GO:0022607)	22.09	2.42E-04	2.96E-02

nucleic acid metabolic process (GO:0090304)	17.6	1.93E-09	1.12E-06
gene expression (GO:0010467)	17.59	9.06E-08	2.80E-05
nucleobase-containing compound metabolic process (GO:0006139)	14.44	7.68E-09	3.96E-06
heterocycle metabolic process (GO:0046483)	13.5	1.23E-08	5.71E-06
cellular aromatic compound metabolic process (GO:0006725)	12.68	1.91E-08	7.38E-06
organic cyclic compound metabolic process (GO:1901360)	12.39	2.25E-08	8.02E-06
cellular nitrogen compound metabolic process (GO:0034641)	10.78	5.96E-08	1.98E-05
macromolecule metabolic process (GO:0043170)	5.01	1.26E-05	2.35E-03
nitrogen compound metabolic process (GO:0006807)	4.64	2.16E-05	3.86E-03
primary metabolic process (GO:0044238)	3.95	6.73E-05	1.16E-02
cellular metabolic process (GO:0044237)	3.88	7.58E-05	1.21E-02
organic substance metabolic process (GO:0071704)	3.75	9.64E-05	1.49E-02
metabolic process (GO:0008152)	3.04	4.14E-04	4.18E-02
Zea mays			
nuclear retention of unspliced pre-mRNA at the site of transcription (GO:0071048)	> 100	3.53E-15	2.37E-12
nuclear mRNA surveillance of spliceosomal pre-mRNA splicing (GO:0071030)	> 100	3.53E-15	2.07E-12
generation of catalytic spliceosome for second transesterification step (GO:0000350)	> 100	3.53E-15	1.84E-12
taxis (GO:0042330)	> 100	5.41E-05	6.21E-03
positive chemotaxis (GO:0050918)	> 100	5.41E-05	6.06E-03
pollen tube guidance (GO:0010183)	> 100	5.41E-05	5.92E-03
chemotaxis (GO:0006935)	> 100	5.41E-05	5.79E-03
intracellular mRNA localization (GO:0008298)	> 100	1.82E-14	8.58E-12
nuclear retention of pre-mRNA at the site of transcription (GO:0071033)	> 100	1.82E-14	7.80E-12
spliceosomal conformational changes to generate catalytic conformation (GO:0000393)	> 100	6.87E-14	2.49E-11
nuclear mRNA surveillance (GO:0071028)	> 100	6.49E-13	1.91E-10
nuclear RNA surveillance (GO:0071027)	> 100	6.49E-13	1.80E-10
RNA surveillance (GO:0071025)	> 100	7.70E-13	1.91E-10
locomotion (GO:0040011)	> 100	2.00E-04	1.81E-02
RNA localization (GO:0006403)	39.52	8.56E-10	1.83E-07
nuclear-transcribed mRNA catabolic process (GO:0000956)	30.55	4.71E-09	9.63E-07
mRNA catabolic process (GO:0006402)	28.91	6.79E-09	1.33E-06
mRNA splicing, via spliceosome (GO:0000398)	28.88	1.31E-15	1.55E-12
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	27.12	2.86E-15	2.69E-12
RNA splicing, via transesterification reactions (GO:0000375)	27.12	2.86E-15	2.24E-12
RNA catabolic process (GO:0006401)	23.54	2.65E-08	4.80E-06
RNA splicing (GO:0008380)	22.47	2.90E-14	1.14E-11
ribonucleoprotein complex assembly (GO:0022618)	17.87	1.63E-07	2.74E-05
ribonucleoprotein complex subunit organization (GO:0071826)	17.58	1.82E-07	2.95E-05

nucleobase-containing compound catabolic process (GO:0034655)	17.44	1.92E-07	3.01E-05
mRNA processing (GO:0006397)	17.35	6.96E-13	1.82E-10
carbohydrate metabolic process (GO:0005975)	17.03	3.43E-36	1.61E-32
cellular nitrogen compound catabolic process (GO:0044270)	15.06	5.02E-07	7.62E-05
heterocycle catabolic process (GO:0046700)	15.01	5.14E-07	7.55E-05
aromatic compound catabolic process (GO:0019439)	13.25	1.16E-06	1.65E-04
defense response (GO:0006952)	13.03	3.51E-08	6.11E-06
organic cyclic compound catabolic process (GO:1901361)	12.82	1.43E-06	1.99E-04
mRNA metabolic process (GO:0016071)	12.5	3.77E-11	8.45E-09
negative regulation of gene expression (GO:0010629)	8.96	1.44E-05	1.82E-03
cellular protein-containing complex assembly (GO:0034622)	7.9	8.25E-06	1.08E-03
RNA processing (GO:0006396)	7.48	1.67E-08	3.13E-06
protein-containing complex assembly (GO:0065003)	7.27	1.49E-05	1.85E-03
negative regulation of macromolecule metabolic process (GO:0010605)	6.41	1.15E-04	1.13E-02
ribonucleoprotein complex biogenesis (GO:0022613)	6.32	1.26E-04	1.21E-02
protein-containing complex subunit organization (GO:0043933)	6.19	4.61E-05	5.42E-03
negative regulation of metabolic process (GO:0009892)	6.15	1.48E-04	1.39E-02
negative regulation of biological process (GO:0048519)	5.65	8.72E-05	8.92E-03
cellular component assembly (GO:0022607)	5.46	1.09E-04	1.09E-02
RNA metabolic process (GO:0016070)	4.35	7.11E-06	9.56E-04
gene expression (GO:0010467)	3.51	3.35E-05	4.04E-03
nucleic acid metabolic process (GO:0090304)	3.04	1.58E-04	1.46E-02
primary metabolic process (GO:0044238)	2.77	4.83E-17	1.14E-13
organic substance metabolic process (GO:0071704)	2.64	4.33E-16	6.79E-13
metabolic process (GO:0008152)	2.21	1.58E-12	3.72E-10
biological_process (GO:0008150)	1.71	2.13E-13	6.69E-11
organonitrogen compound metabolic process (GO:1901564)	0.09	6.33E-05	6.62E-03
Gossypium hirsutum			
negative regulation of ATPase activity (GO:0032780)	> 100	2.40E-07	8.29E-05
negative regulation of DNA duplex unwinding (GO:1905463)	> 100	2.40E-07	7.65E-05
regulation of DNA duplex unwinding (GO:1905462)	> 100	2.40E-07	7.10E-05
negative regulation of helicase activity (GO:0051097)	> 100	2.40E-07	6.63E-05
generation of catalytic spliceosome for second transesterification step (GO:0000350)	> 100	2.40E-07	6.21E-05
negative regulation of DNA helicase activity (GO:1905775)	> 100	2.40E-07	5.85E-05
regulation of DNA helicase activity (GO:1905774)	> 100	2.40E-07	5.52E-05
nuclear retention of unspliced pre-mRNA at the site of transcription (GO:0071048)	> 100	1.50E-06	2.59E-04
nuclear mRNA surveillance of spliceosomal pre-mRNA splicing (GO:0071030)	> 100	1.50E-06	2.48E-04
cell cycle DNA replication initiation (GO:1902292)	> 100	2.47E-06	3.79E-04
spliceosomal conformational changes to generate catalytic conformation (GO:0000393)	> 100	2.47E-06	3.66E-04

nuclear cell cycle DNA replication initiation (GO:1902315)	> 100	2.47E-06	3.53E-04
mitotic DNA replication initiation (GO:1902975)	> 100	2.47E-06	3.41E-04
oligosaccharide transport (GO:0015772)	97.47	6.53E-06	8.20E-04
sucrose transport (GO:0015770)	97.47	6.53E-06	7.96E-04
disaccharide transport (GO:0015766)	97.47	6.53E-06	7.73E-04
nuclear retention of pre-mRNA at the site of transcription (GO:0071033)	91.74	7.68E-06	8.37E-04
intracellular mRNA localization (GO:0008298)	91.74	7.68E-06	8.15E-04
regulation of helicase activity (GO:0051095)	86.64	8.94E-06	9.26E-04
pre-replicative complex assembly involved in cell cycle DNA replication (GO:1902299)	74.27	1.36E-05	1.34E-03
pre-replicative complex assembly (GO:0036388)	74.27	1.36E-05	1.31E-03
pre-replicative complex assembly involved in nuclear cell cycle DNA replication (GO:0006267)	74.27	1.36E-05	1.28E-03
mitotic DNA replication (GO:1902969)	64.98	1.95E-05	1.76E-03
DNA unwinding involved in DNA replication (GO:0006268)	59.98	2.43E-05	2.14E-03
negative regulation of chromosome organization (GO:2001251)	53.78	3.28E-05	2.78E-03
double-strand break repair via break-induced replication (GO:0000727)	51.99	3.61E-05	2.99E-03
nuclear mRNA surveillance (GO:0071028)	45.87	5.11E-05	4.15E-03
nuclear RNA surveillance (GO:0071027)	45.87	5.11E-05	4.07E-03
RNA surveillance (GO:0071025)	42.15	6.47E-05	5.06E-03
defense response (GO:0006952)	39.25	1.61E-63	6.67E-60
nuclear DNA replication (GO:0033260)	38.04	8.62E-05	6.61E-03
defense response to other organism (GO:0098542)	36.89	1.64E-27	2.26E-24
response to other organism (GO:0051707)	36.66	1.87E-27	1.94E-24
response to external biotic stimulus (GO:0043207)	36.66	1.87E-27	1.55E-24
interspecies interaction between organisms (GO:0044419)	34.24	7.71E-27	4.56E-24
cell cycle DNA replication (GO:0044786)	33.9	1.19E-04	8.81E-03
DNA replication initiation (GO:0006270)	33.9	1.19E-04	8.66E-03
response to biotic stimulus (GO:0009607)	29.48	1.74E-25	8.99E-23
regulation of ATPase activity (GO:0043462)	24.76	2.89E-04	2.03E-02
sucrose metabolic process (GO:0005985)	23.28	3.44E-04	2.37E-02
response to external stimulus (GO:0009605)	19.09	1.43E-21	6.57E-19
regulation of chromosome organization (GO:0033044)	19.02	6.07E-04	3.99E-02
negative regulation of organelle organization (GO:0010639)	18.13	6.94E-04	4.43E-02
response to stress (GO:0006950)	12.08	2.17E-42	4.50E-39
mRNA splicing, via spliceosome (GO:0000398)	11.39	6.96E-07	1.52E-04
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	11.21	7.84E-07	1.63E-04
RNA splicing, via transesterification reactions (GO:0000375)	11.21	7.84E-07	1.55E-04
RNA splicing (GO:0008380)	9.39	2.85E-06	3.81E-04
mRNA processing (GO:0006397)	6.9	2.55E-05	2.20E-03
response to stimulus (GO:0050896)	5.66	3.12E-27	2.16E-24
oxidation-reduction process (GO:0055114)	5.57	3.29E-20	1.36E-17
mRNA metabolic process (GO:0016071)	5.08	2.07E-04	1.48E-02

metabolic process (GO:0008152)	1.43	6.76E-04	4.37E-02
biological_process (GO:0008150)	1.37	1.08E-06	1.95E-04
cellular metabolic process (GO:0044237)	0.53	7.09E-04	4.45E-02
cellular process (GO:0009987)	0.4	7.70E-09	2.90E-06
cellular protein metabolic process (GO:0044267)	0.26	5.42E-04	3.62E-02
protein metabolic process (GO:0019538)	0.23	1.16E-04	8.74E-03
organonitrogen compound metabolic process (GO:1901564)	0.19	4.37E-06	5.65E-04
cellular protein modification process (GO:0006464)	< 0.01	7.64E-06	8.79E-04
macromolecule modification (GO:0043412)	< 0.01	1.96E-06	3.13E-04
protein modification process (GO:0036211)	< 0.01	7.64E-06	8.56E-04
phosphate-containing compound metabolic process (GO:0006796)	< 0.01	1.93E-05	1.78E-03
phosphorus metabolic process (GO:0006793)	< 0.01	1.18E-05	1.19E-03
phosphorylation (GO:0016310)	< 0.01	3.80E-04	2.58E-02

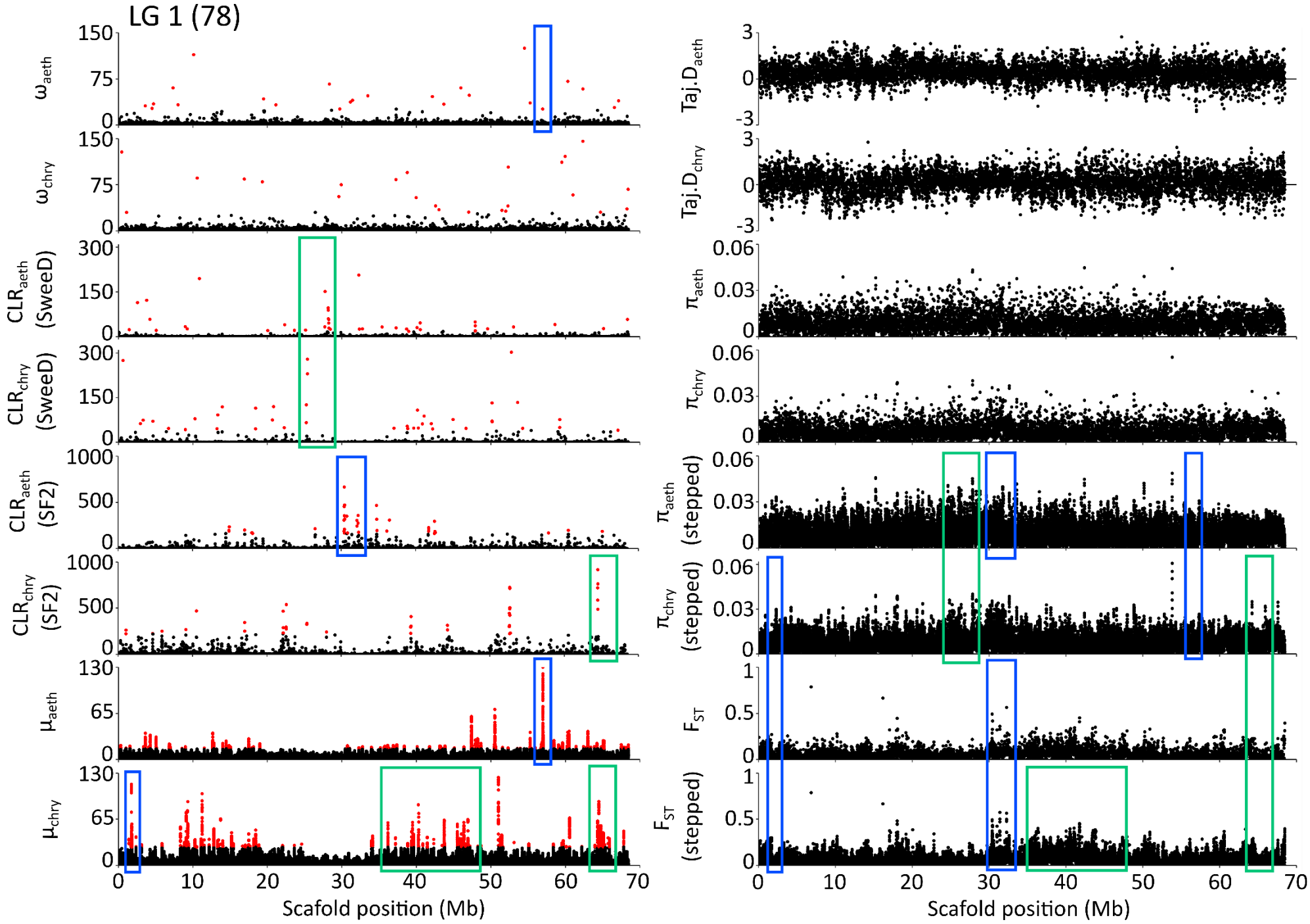
Table 4.10. List of over-represented GO terms for molecular functions using genes identified in Table 4.8 and annotations from two species (*Zea mays*, *Gossypium hirsutum*). Analysis using *Helianthus annus* did not come back with any over-represented GO terms.

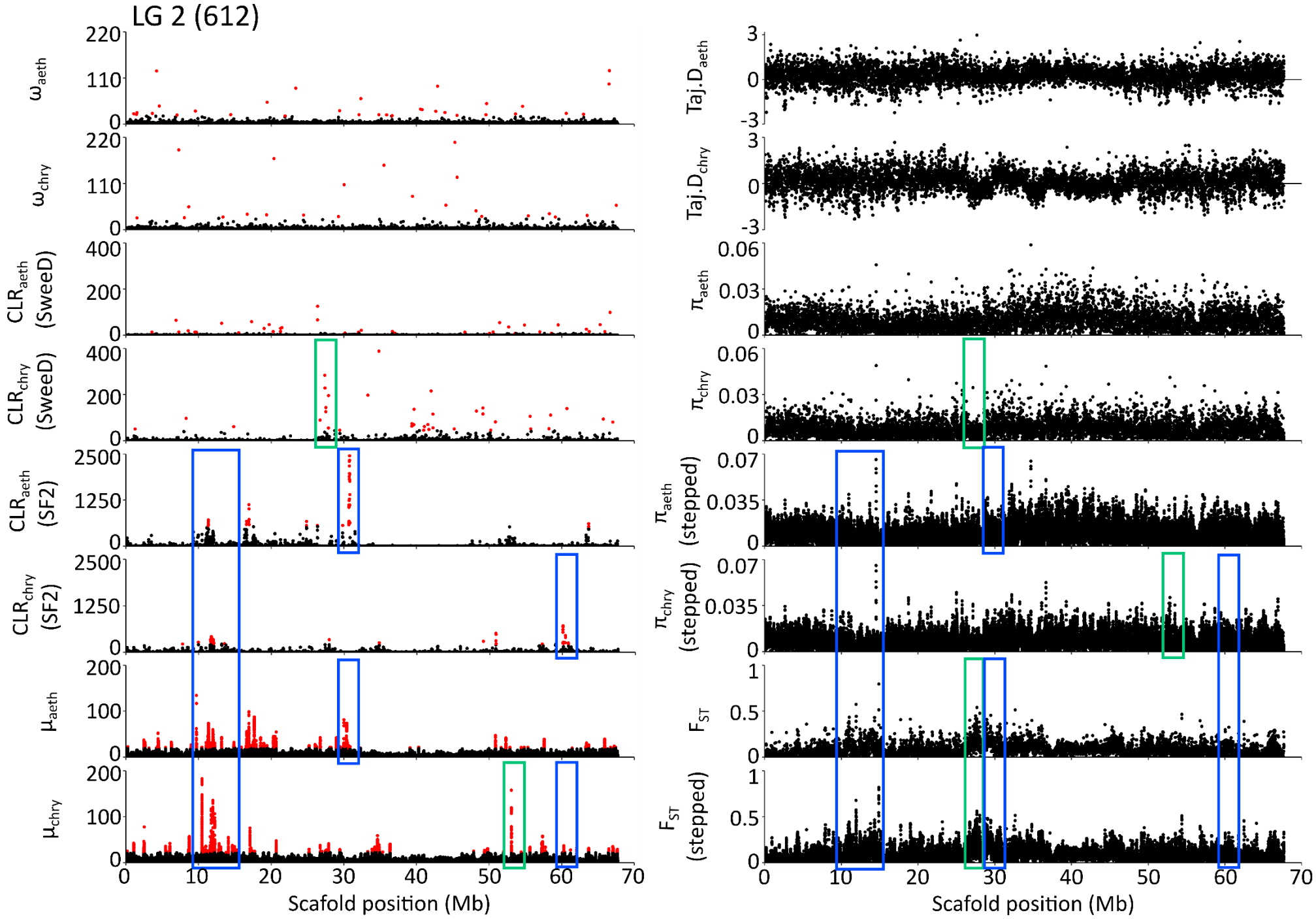
	Fold enrichment	Raw P-value	FDR P-value
Zea mays			
mediator complex binding (GO:0036033)	> 100	1.55E-05	3.94E-03
polygalacturonase activity (GO:0004650)	> 100	1.09E-80	2.76E-77
hydrolase activity, hydrolyzing O-glycosyl compounds (GO:0004553)	47.54	1.33E-51	1.69E-48
hydrolase activity, acting on glycosyl bonds (GO:0016798)	42.2	8.02E-50	6.79E-47
lyase activity (GO:0016829)	30.07	4.72E-34	3.00E-31
hydrolase activity (GO:0016787)	4.89	4.11E-19	2.09E-16
catalytic activity (GO:0003824)	1.79	8.73E-06	2.46E-03
binding (GO:0005488)	0.18	1.03E-07	4.36E-05
ion binding (GO:0043167)	0.16	2.10E-04	4.85E-02
heterocyclic compound binding (GO:1901363)	0.13	3.81E-06	1.21E-03
organic cyclic compound binding (GO:0097159)	0.12	3.78E-06	1.37E-03
Gossypium hirsutum			
sucrose:cation symporter activity (GO:0009669)	97.47	6.53E-06	1.43E-03
sucrose transmembrane transporter activity (GO:0008515)	97.47	6.53E-06	1.31E-03
sucrose:proton symporter activity (GO:0008506)	97.47	6.53E-06	1.21E-03
single-stranded DNA helicase activity (GO:0017116)	97.47	6.53E-06	1.12E-03
carbohydrate:cation symporter activity (GO:0005402)	77.98	1.19E-05	1.90E-03
oligosaccharide transmembrane transporter activity (GO:0015157)	77.98	1.19E-05	1.78E-03
disaccharide transmembrane transporter activity (GO:0015154)	77.98	1.19E-05	1.68E-03
carbohydrate:proton symporter activity (GO:0005351)	77.98	1.19E-05	1.59E-03
solute:proton symporter activity (GO:0015295)	42.15	6.47E-05	8.18E-03
solute:cation symporter activity (GO:0015294)	38.04	8.62E-05	1.04E-02
DNA replication origin binding (GO:0003688)	37.13	9.22E-05	1.06E-02
3'-5' DNA helicase activity (GO:0043138)	36.27	9.85E-05	1.08E-02
monooxygenase activity (GO:0004497)	27.48	7.06E-42	3.40E-39
iron ion binding (GO:0005506)	26.48	5.71E-46	1.37E-42
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen (GO:0016705)	23.53	5.72E-44	6.87E-41
heme binding (GO:0020037)	23.45	6.50E-44	5.21E-41
tetrapyrrole binding (GO:0046906)	21.93	8.84E-43	5.31E-40
transition metal ion binding (GO:0046914)	8.36	8.32E-29	3.33E-26
oxidoreductase activity (GO:0016491)	5.99	2.29E-21	7.88E-19
metal ion binding (GO:0046872)	3.74	1.65E-16	4.97E-14
cation binding (GO:0043169)	3.71	2.27E-16	6.07E-14
ion binding (GO:0043167)	1.91	2.23E-06	5.35E-04
transferase activity (GO:0016740)	0.28	2.62E-04	2.63E-02
transferase activity, transferring phosphorus-containing groups (GO:0016772)	< 0.01	1.65E-04	1.72E-02

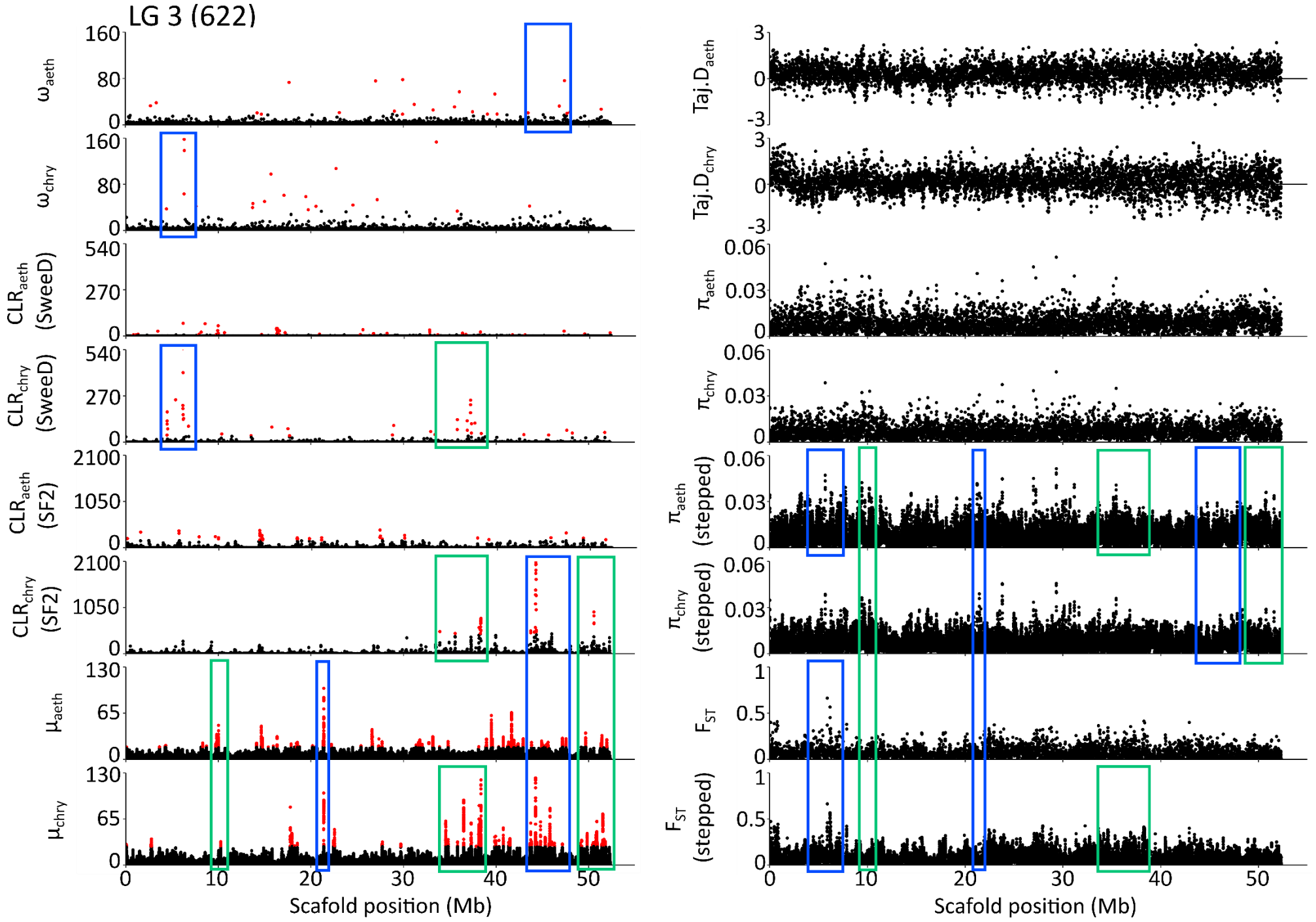
Table 4.11. List of over-represented GO terms for cellular processes using genes identified in Table 4.8 and annotations from three species (*Helianthus annus*, *Zea mays*, *Gossypium hirsutum*).

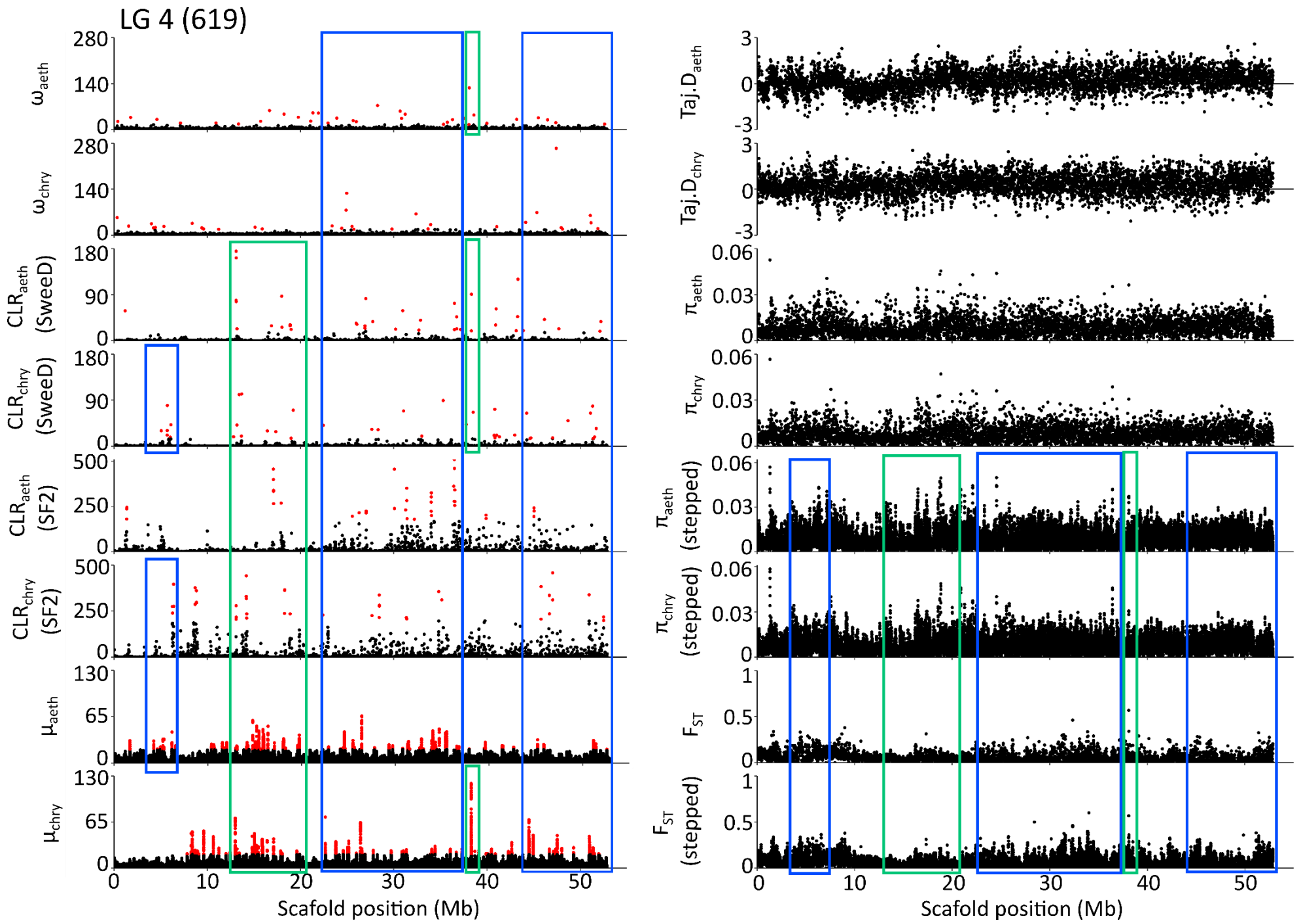
	Fold enrichment	Raw P-value	FDR P-value
Helianthus annus			
U2-type post-spliceosomal complex (GO:0071021)	> 100	9.67E-08	1.80E-05
post-spliceosomal complex (GO:0071020)	> 100	1.61E-07	2.50E-05
precatalytic spliceosome (GO:0071011)	> 100	1.52E-11	7.08E-09
U5 snRNP (GO:0005682)	> 100	4.44E-06	5.16E-04
U4/U6 x U5 tri-snRNP complex (GO:0046540)	> 100	1.38E-05	1.43E-03
spliceosomal tri-snRNP complex (GO:0097526)	> 100	2.13E-05	1.98E-03
U2-type spliceosomal complex (GO:0005684)	> 100	4.33E-05	3.36E-03
spliceosomal complex (GO:0005681)	> 100	6.90E-14	6.41E-11
spliceosomal snRNP complex (GO:0097525)	> 100	6.28E-05	4.17E-03
small nuclear ribonucleoprotein complex (GO:0030532)	> 100	6.28E-05	3.89E-03
Sm-like protein family complex (GO:0120114)	> 100	8.09E-05	4.17E-03
ribonucleoprotein complex (GO:1990904)	41.38	5.53E-10	1.71E-07
protein-containing complex (GO:0032991)	11.81	3.13E-08	7.28E-06
nucleus (GO:0005634)	7.85	5.49E-07	7.28E-05
intracellular membrane-bounded organelle (GO:0043231)	4.23	4.10E-05	3.46E-03
membrane-bounded organelle (GO:0043227)	4.18	4.49E-05	3.21E-03
intracellular organelle (GO:0043229)	3.87	7.76E-05	4.50E-03
organelle (GO:0043226)	3.86	7.79E-05	4.25E-03
intracellular anatomical structure (GO:0005622)	3.2	2.89E-04	1.41E-02
Zea mays			
U2-type post-spliceosomal complex (GO:0071021)	> 100	7.17E-17	1.35E-14
post-spliceosomal complex (GO:0071020)	> 100	1.61E-15	1.68E-13
U5 snRNP (GO:0005682)	> 100	2.09E-13	1.97E-11
U4/U6 x U5 tri-snRNP complex (GO:0046540)	93.65	3.06E-12	2.62E-10
precatalytic spliceosome (GO:0071011)	78.56	3.10E-10	2.09E-08
spliceosomal tri-snRNP complex (GO:0097526)	73.01	1.52E-11	1.20E-09
cell wall (GO:0005618)	63.12	8.33E-56	7.84E-53
external encapsulating structure (GO:0030312)	62.76	1.01E-55	4.76E-53
U2-type spliceosomal complex (GO:0005684)	56.68	7.95E-11	5.76E-09
spliceosomal snRNP complex (GO:0097525)	43.08	4.84E-10	3.04E-08
small nuclear ribonucleoprotein complex (GO:0030532)	42.65	5.17E-10	3.04E-08
Sm-like protein family complex (GO:0120114)	40.64	7.12E-10	3.94E-08
spliceosomal complex (GO:0005681)	37.74	4.84E-17	1.14E-14
cell periphery (GO:0071944)	9.17	5.77E-27	1.81E-24
ribonucleoprotein complex (GO:1990904)	7.83	9.87E-09	5.16E-07
extracellular region (GO:0005576)	6.93	6.11E-06	3.03E-04
cellular anatomical entity (GO:0110165)	1.81	8.99E-16	1.41E-13
cellular_component (GO:0005575)	1.8	1.01E-15	1.36E-13
membrane (GO:0016020)	0.31	7.34E-04	3.29E-02
cytoplasm (GO:0005737)	0.26	3.36E-04	1.58E-02
Gossypium hirsutum			
U2-type post-spliceosomal complex (GO:0071021)	> 100	1.37E-07	2.47E-05
post-spliceosomal complex (GO:0071020)	> 100	2.40E-07	3.59E-05

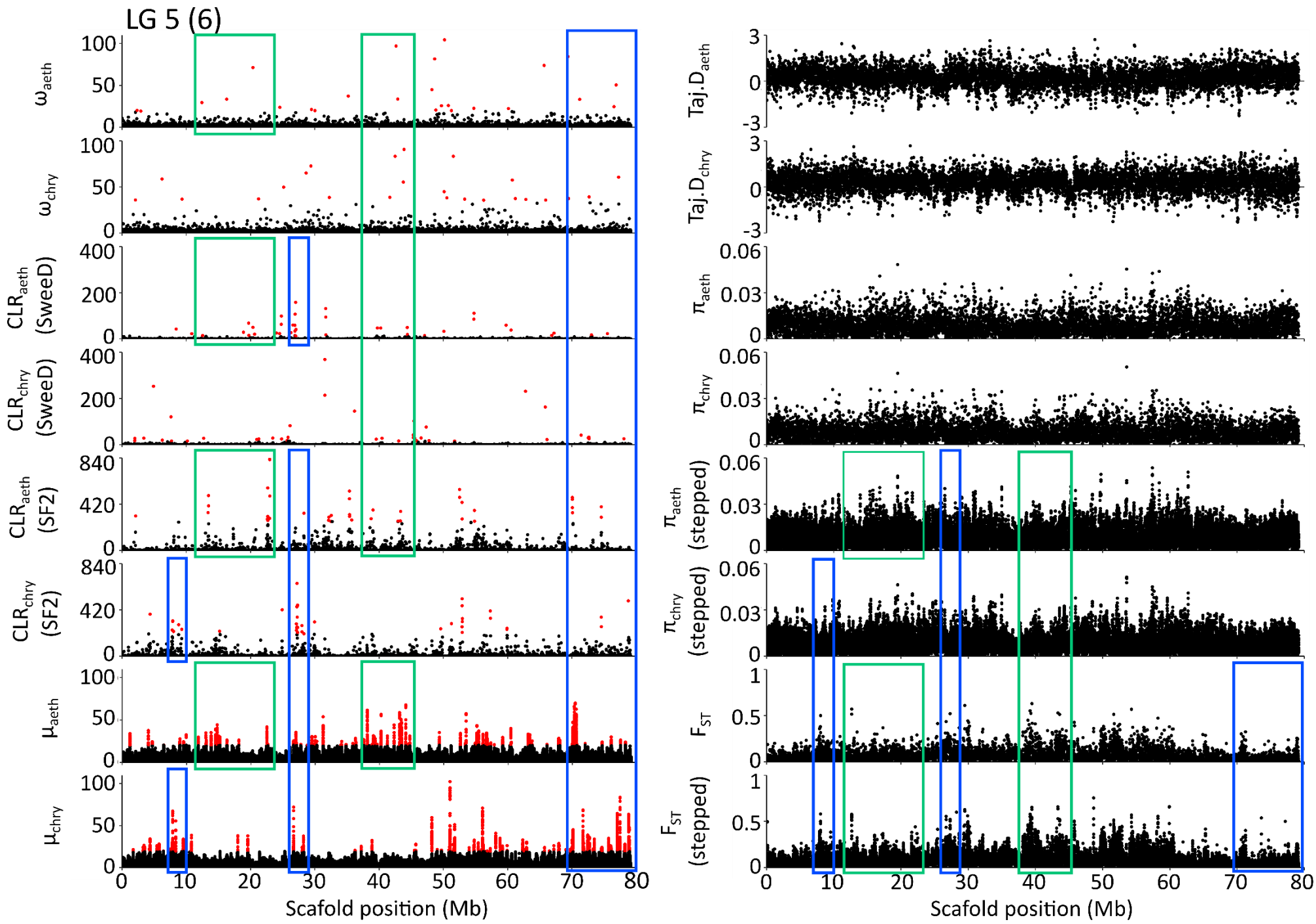
MCM complex (GO:0042555)	67.81	1.74E-05	1.95E-03
precatalytic spliceosome (GO:0071011)	53.05	7.01E-08	3.15E-05
U5 snRNP (GO:0005682)	42.15	6.47E-05	6.45E-03
U4/U6 x U5 tri-snRNP complex (GO:0046540)	36.27	9.85E-05	8.85E-03
spliceosomal tri-snRNP complex (GO:0097526)	23.28	3.44E-04	2.81E-02
spliceosomal complex (GO:0005681)	14.91	9.58E-08	2.15E-05
extracellular region (GO:0005576)	11.44	7.53E-24	6.76E-21
intracellular anatomical structure (GO:0005622)	0.46	4.42E-06	5.67E-04
cytoplasm (GO:0005737)	0.19	8.38E-08	2.51E-05

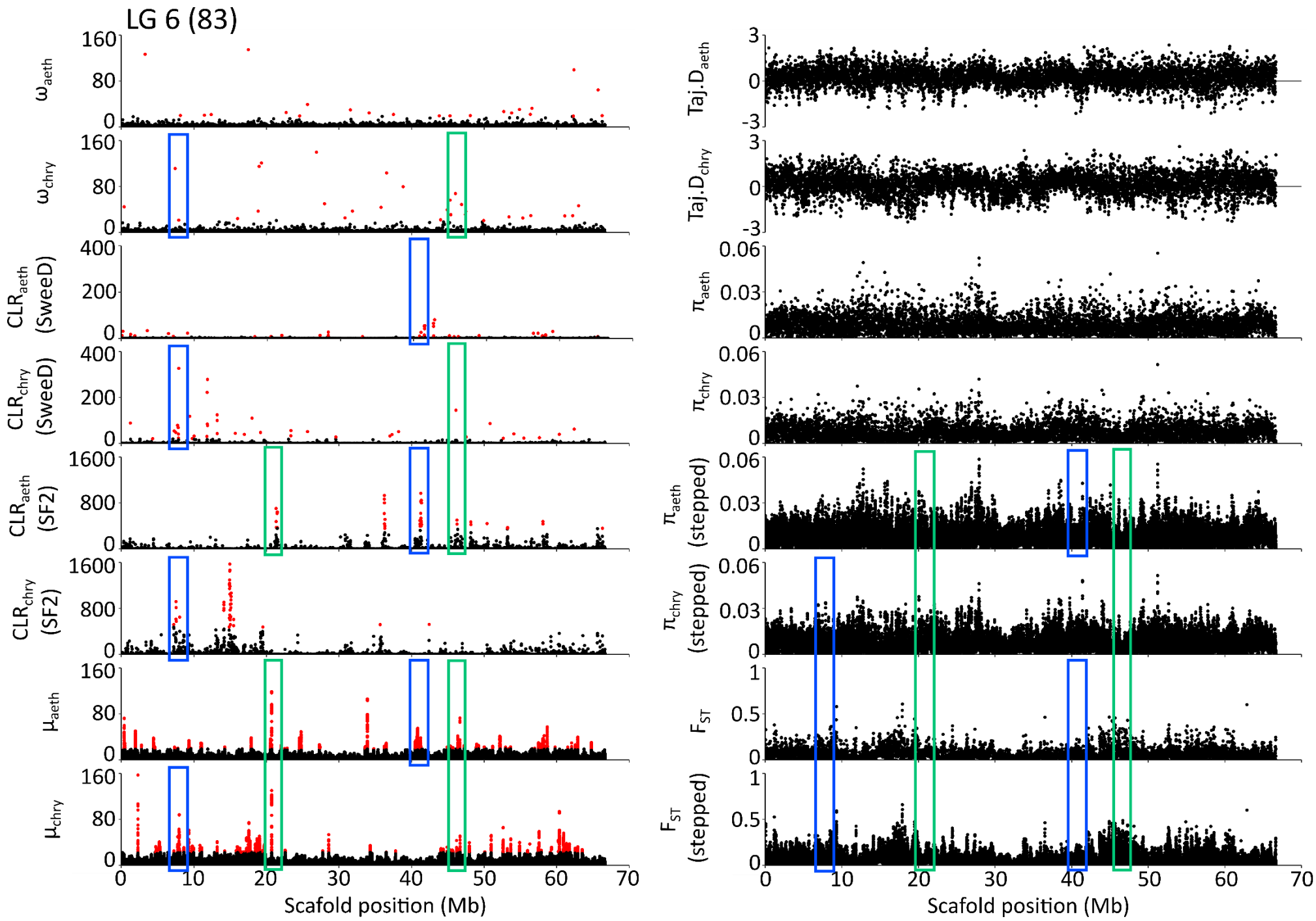

LG 1 (78)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)

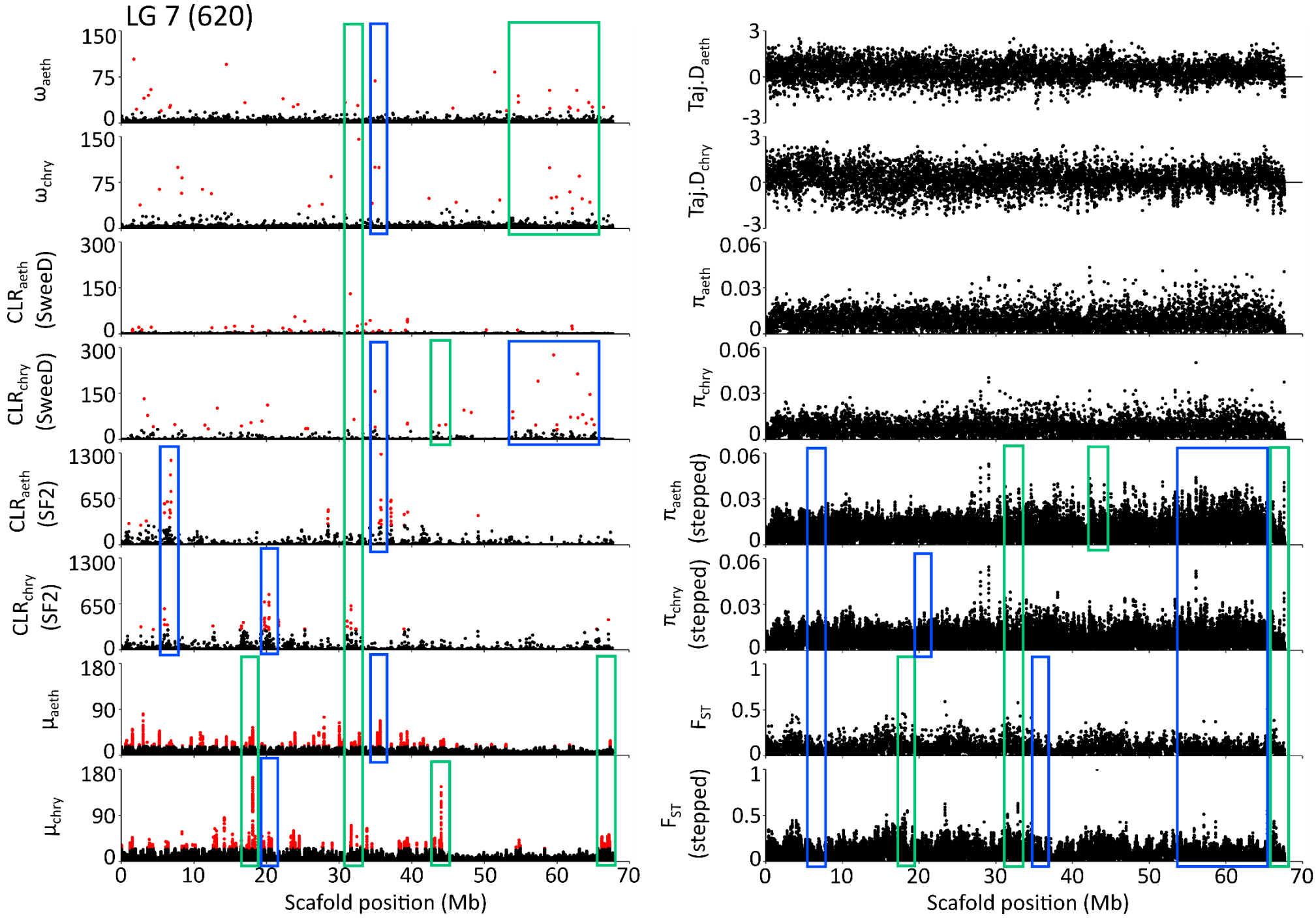

LG 2 (612)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)

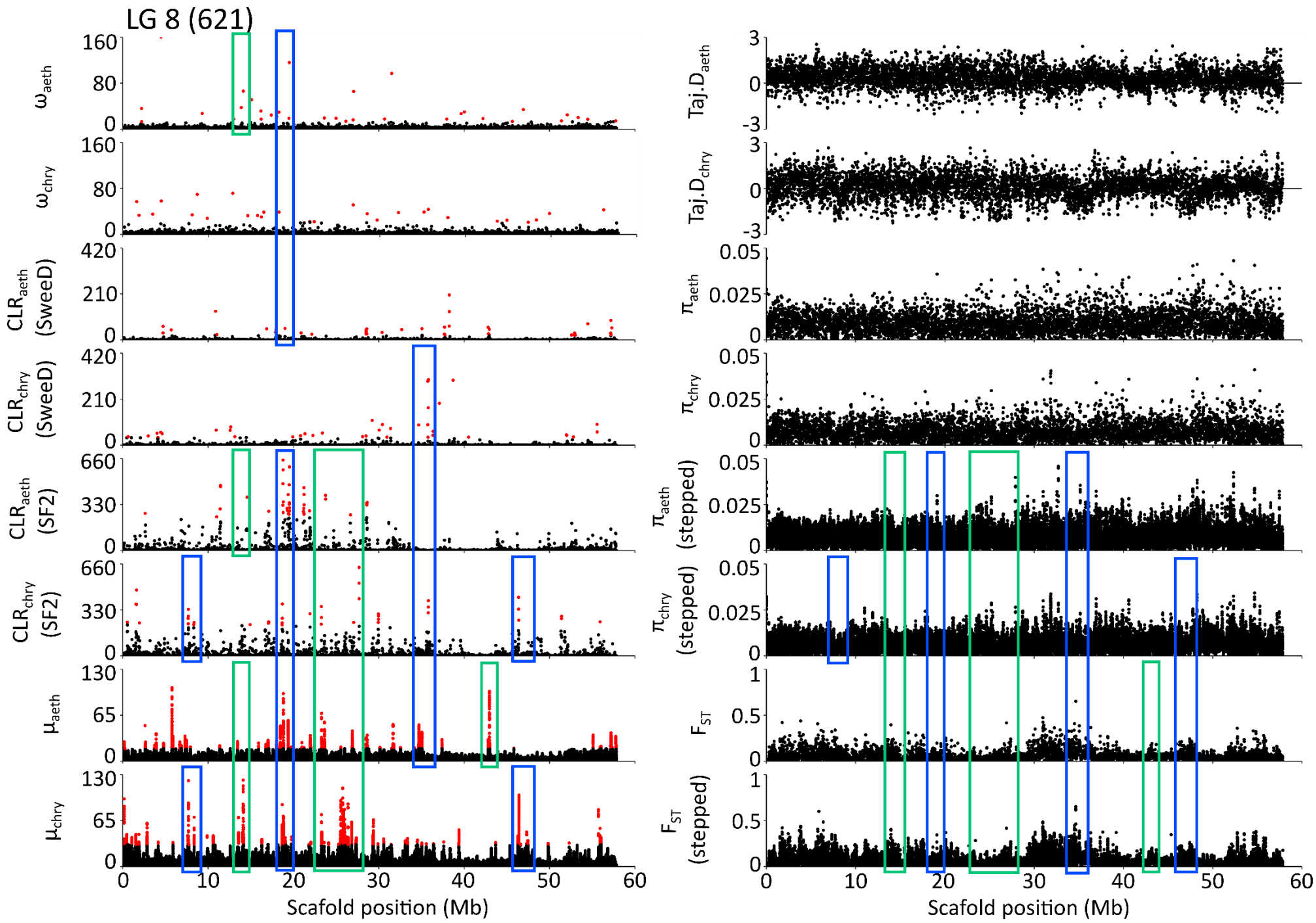

LG 3 (622)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)

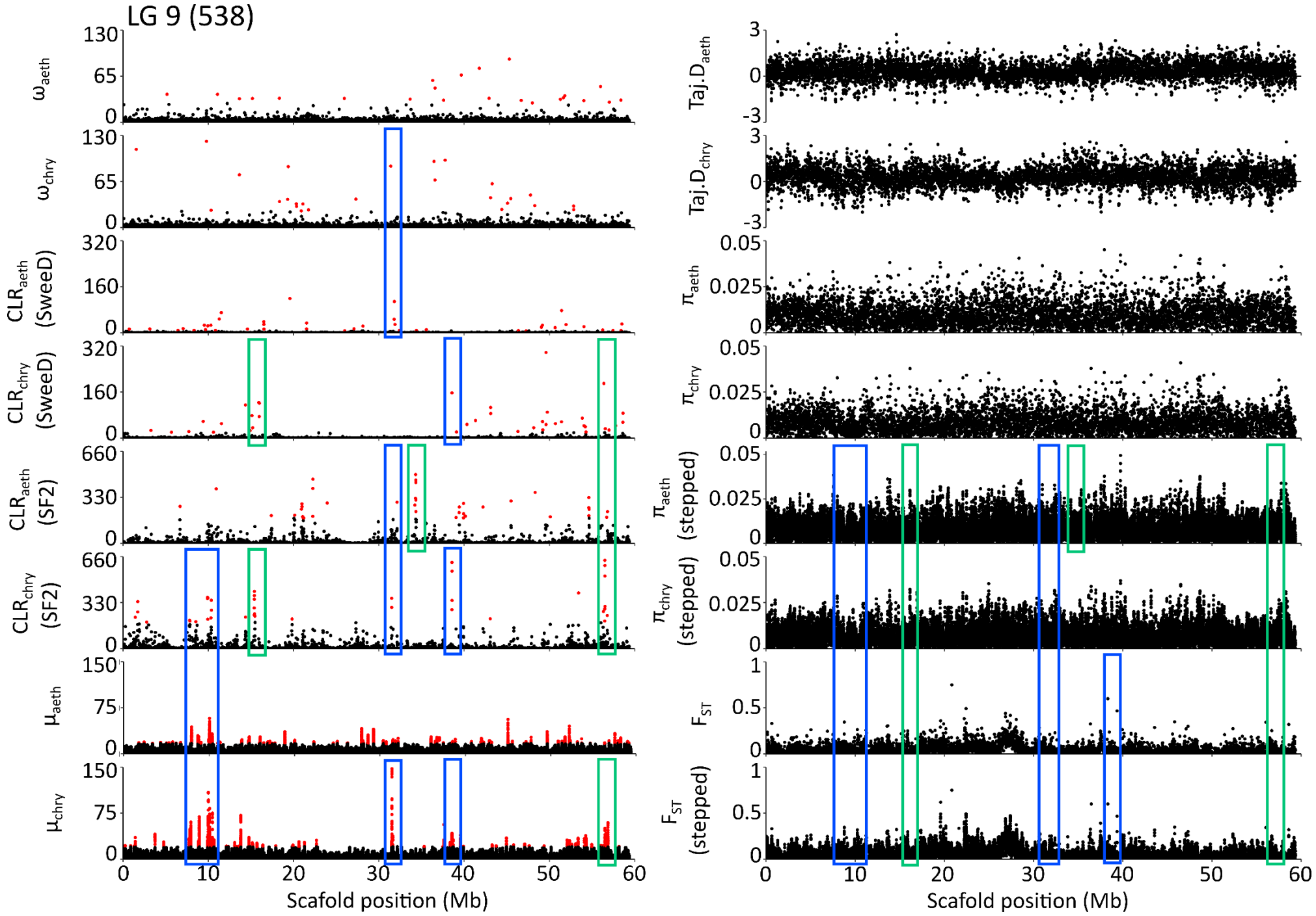

LG 4 (619)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)


LG 5 (6)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)


LG 6 (83)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)


LG 7 (620)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)


LG 8 (621)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)


LG 9 (538)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)

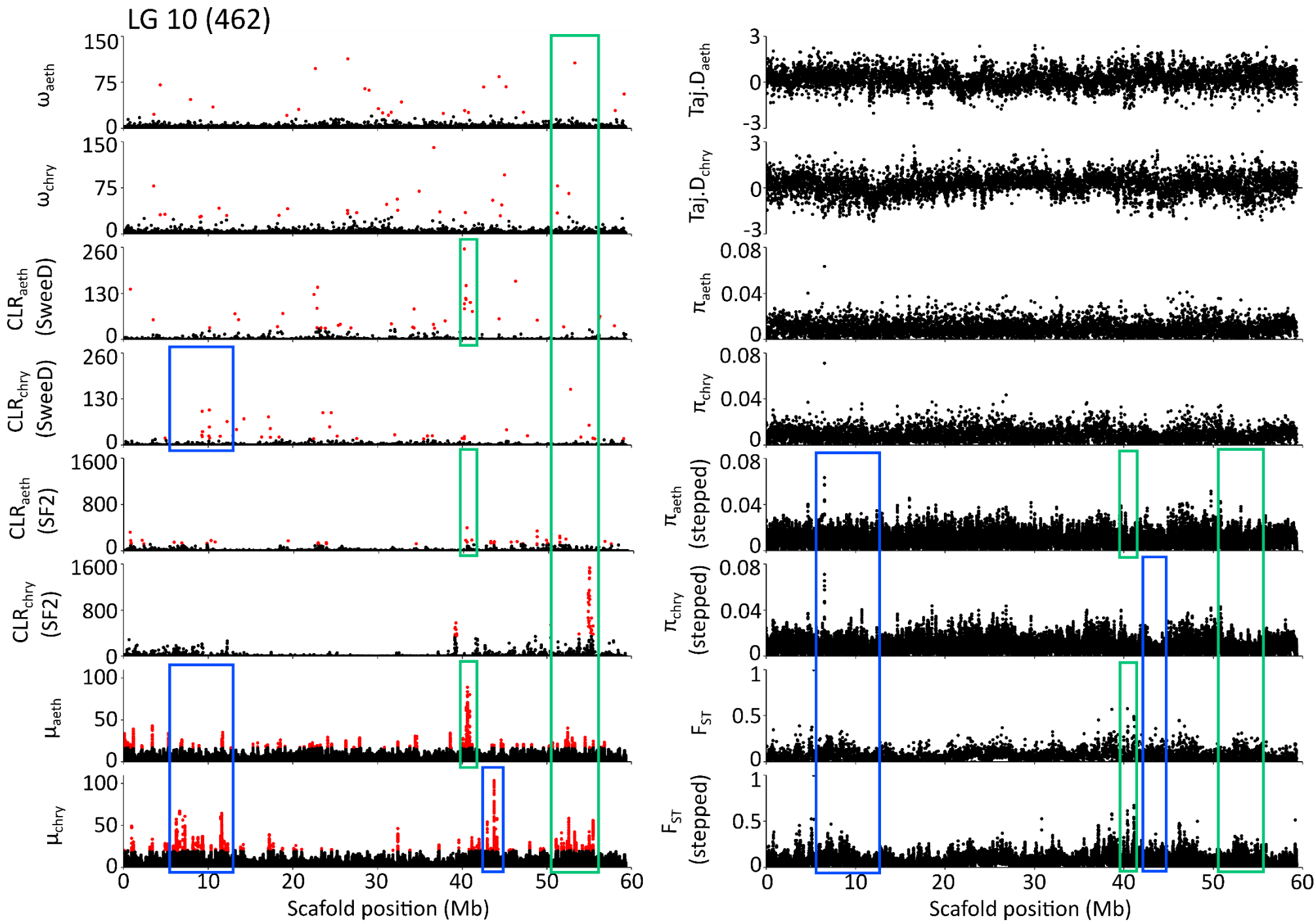

LG 10 (462)
ωaeth
ωchry
CLRaeth (SweeD)
CLRchry (SweeD)
CLRaeth (SF2)
CLRchry (SF2)
μaeth
μchry
Taj.Daeth
Taj.Dchry
πaeth
πchry
πaeth (stepped)
πchry (stepped)
FST
FST (stepped)
Scafold position (Mb)
Scafold position (Mb)

Figure 4.3. Plots for selective sweeps scans and polymorphisms along each scaffold. For each linkage group (genomic scaffold), sweep scans are on the left and polymorphism plots are on the right. From top to bottom of the sweep scans, ω statistics are outputs from *OmegaPlus*, CLRs are from *SweeD* and *SweepFinder2* and μ is from *RAiSD*. All polymorphisms are estimated using 10,000b sliding windows; step size is 2,000b for overlapping sliding windows. Regions with high or clusters of sweep signals were boxed alternatively with blue and green rectangles, with corresponding coloured regions in the polymorphism plots. Analyses and genes matched to these regions can be found in Table 4.3.

4.5 Discussion

In this study, I have created an integrated map using the *S. squalidus* reference genome and markers from the linkage map of Chapman *et al.* (2013). I then used various markers to study the relationship between recombination and nucleotide diversity in the ten linkage groups. With the genome sequences of *S. aethnensis* and *S. chrysanthemifolius*, genome wide polymorphism patterns were estimated and genome scans for signals of selective sweeps were carried out, yielding multiple regions with putative sweeps within and across linkage groups. The putative sweeps were estimated to have various strengths and ages. They also harboured many genes with functions and processes potentially related to adaptation to high and low elevations on Mount Etna.

4.5.1 Correlation between recombination rate and DNA polymorphism

Chromosomal and genomic structure were not well-studied in the Etnean *Senecio* system. The integrated maps reported here provided a basic idea of the former – linkage groups with high recombination at only one side of the genomic scaffold potentially means the centromere is skewed towards one side (LG 3, 7). The observation that π was significantly lower in low-recombining regions in all linkage groups in *S. aethnensis* also fits theory that this positive correlation may be a result of fixation of advantageous mutations (positive selection) and genetic hitchhiking (Figure 4.1; Begun & Aquadro, 1992). However, this was not the case for five of the linkage groups of *S. chrysanthemifolius*. When the correlation of local recombination rates and π was tested with the three groups of nextRAD and differentially expressed markers, the positive correlation was only significant for markers in four and two linkage groups for *S. aethnensis* and *S. chrysanthemifolius* respectively (others also showed correlation but did not show significance). Linkage group 3 and 10 were the only two which showed a significant positive correlation between recombination rates and π in both species, suggesting that positive selection in them could be more prominent than that in other linkage groups. As software tools detected signals of selective sweeps in all linkage groups (discussed below), the lack of significant correlation between recombination rates and π in some linkage groups is likely due to a lack of data

from markers in low-recombining regions (most markers were located in medium- to high-recombining regions as shown in Figure 4.1). In theory, the correlation is driven by a lack of recombination in regions, resulting in strong genetic hitchhiking and loss of diversity (low π). Markers in medium- to high-recombining regions would have similar π (hence no correlation with recombination rate) as recombination, even a little, would prevent genetic hitchhiking so diversity is maintained. Estimation of recombination rates is also constrained by the number of markers in different linkage positions. In this study, many markers in the linkage map occupied the same position (cM) hence most of them were unsuitable for use (as there would be 'zero' linkage map distance between those markers). This linkage map regions with low resolution would then lead to inaccurate calculation of recombination rate as it does not reflect the true location of markers – for instance, markers that are at the same cM on the linkage map might be tens of thousands of bases away from each other in the genome. Denser linkage mapping with more markers and resolution may be able to provide clarity in testing the relationship between recombination and π again in the future.

4.5.2 Signals of putative selective sweeps and patterns of polymorphism

V-shape valleys have been observed in the π polymorphism plots in this study for various sweeps (Figure 4.3). However, other patterns of nucleotide diversity and different patterns of F_{ST} were present as well. Taking sweep regions (a) 4 -7Mb (for *S. chrysanthemifolius*), (b) 9 – 11Mb (for both species) and (c) 20 – 22Mb (for both species) of LG 3 as examples (Table 4.8), (b) showed a V-shape valley for π in both species. (a) also showed a V-shape valley in π for *S. chrysanthemifolius* where there were sweep signals, but there was a more prominent V-shape valley in π for *S. aethnensis* albeit software tools did not detect any significant sweep signals in the region. Instead of having a V-shape valley or low π, (c) seemingly had a high peak for π in both species where the sweep signal was. The high peaks for π were likely not correspondent to sweep signals but rather other processes near the sweeps. Compared to π, F_{ST} in these regions better fitted theoretical expectations. In (a), as there were only sweep signals in *S. chrysanthemifolius*, F_{ST} was high. F_{ST} in (b) was low as there were sweep signals

in both species; likewise, (c) had reduced F_{ST} in the region, creating a V-shape valley similar to those in π. On the contrary, F_{ST} in regions where both species showed sweep signals was high in other scaffolds, such as 10 – 15Mb in LG 2. This could be due to different mutations were being driven to fixation in each of the species hence creating differentiation between species [while the same or similar mutations were being fixed in (b) above hence population differentiation was low]. The resolution of the graphs could affect the interpretations of the polymorphism patterns. However, given hundreds of regions showed strong sweep signals within and across scaffolds, it would be challenging to examine all sweep signals without clear targets of selection; compared with only 22 selective sweeps identified in African highland-adapted maize, *Zea mays ssp. mays*: Wegary *et al.*, 2019; and only two sweeps identified in winter-grey mountain hares, *Lepus timidus*, one of which corresponded to the pigmentation gene *Agouti*: Giska *et al.*, 2019. The former study used SweeD to detect sweeps in 298 samples, and considered regions outliers if they were in the top 0.1% of sweep signals; whereas the latter study used a Population Branch Statistic (PBS) to detect sweeps in 59 samples, and considered regions outliers for those in the top 0.01% of sweep signals. Although these and the current studies used different number of samples, softwares and cut-off percentage to identify outliers, what is obvious is that unlike other studies which had clear peaks of outliers that were localised, that was not the case in this study and there would still be multiple targets of sweeps widespread within and across each linkage group if similar criteria were used to detect sweeps. There was also a lack of common outliers between the two species in the same software tool and between software tools in each species. The former might suggest that each species was positively selected for different genes, which is not surprising as they are adapted to different environments. For the latter, the lack of outliers in common for *OmegaPlus* and the other two software tools is possibly due to that fact that *OmegaPlus* identifies sweeps based on LD, while the other two identify sweeps based on SFS. However, *SweeD* and *SweepFinder2*, which both utilises SFS, also shared very few outliers. This could be due to different algorithms in each software tools.

4.5.3 False-positives in genome scans

Out of the 57 regions with sweeps signals with polymorphism patterns and corresponding markers analysed, only 32 had high confidence of being true sweeps based on their polymorphism patterns. The rest were likely false-positives or require more in-depth analyses (there could also be false-negatives not being detected). Besides, *SweeD* and *SweepFinder2* did not share most of their sweep outliers even though they were both based on SFS. Together with the abundance of high peaks in *RAiSD*, these are also suggestive of the presence of false-positives. False-positives could be due to demographic processes which generate signals that resembles those of selective sweeps, or inaccuracies in sequencing (as mentioned in the *Introduction*). Hence, the results should be interpreted with caution, especially since population size change and other demographic processes were inferred in the system in *Chapter 3* (Wong *et al.*, 2020). One way to reduce false-positives is to use the approach employed in this study - to look for overlapping signals across multiple software tools. This is based on the assumption that signals from more than one software tools would give a higher confidence on the possibility of the sweep signals being true positives. However, this approach still resulted in many candidate regions and genes with functions that are related to elevation and volcanic substrates on Mount Etna. Therefore, it is crucial that promising candidates are validated through future independent re-sequencing projects, as advocated by studies such as Mallick *et al.* (2009).

4.5.4 Strength and age of selective sweeps

Comparing the outputs from *SweeD* and *SweepFinder2*, it is clear that the two software tools provided quite different results for sweeps at the same location – *SweepFinder2* yielded α, and subsequently selection coefficient (*s)* and age of sweeps (T) estimates that were multiples of those from *SweeD* in sweeps at the same location. For instance, all *s* estimates from *SweeD* were below 0.14 whereas those from *SweepFinder2* were as high as 1.08. *s* larger than one is likely due to the recombination rate used in estimations being overestimated, as it was based on the limited number of markers used in this study. If a lower recombination rate was used, *s* would be smaller and T would

be larger across all the estimations. Together with the fact that estimation of the *s* and T were also based on effective population sizes and generation time, results should be interpreted with caution.

Age of sweeps estimated were all fewer than 1250 generations (Table 4.7). Given the generation time of one to a few years in Etnean *Senecio* (Brennan *et al.*, 2009), the ages of the selective sweeps were all within a few thousand years old (assuming α estimates from software tools are accurate, these age estimates are unusually recent likely due to imprecise recombination rate used as explained above). No massive climatic events had happened in this period, thus it is unlikely the selective sweeps were caused by them. A possible explanation for the recent sweeps could be recent geological events such as eruptions of Mount Etna that led to a change in environmental conditions (for example, shifts in flora, fauna and soil compositions) and hence new selective pressures. For sweeps that were only tens of generations old, recent increase in human activities (tourism and the wine industry) could be imposing selective pressures responsible for them. This is in line with Ross (2010), who proposed that human activities had aided dispersal and gene flow between *S. aethnensis* and *S. chrysanthemifolius*. As mentioned above, recombination rate used in this study was likely overestimated hence it is likely that sweeps presented would have older ages.

This analysis also showed that some common sweeps between *S. aethnensis* and *S. chrysanthemifolius* had the same strength and age, while others did not. Common sweep age could indicate that divergent selection was the main form of selection in those selective sweeps. Similar to findings on multifarious divergent selection in *Chapter 3* (Wong *et al.*, 2020), this study also shows that selective sweeps were in various strengths and ages.

4.5.5 Genes in genomic regions with putative sweeps and their implications for adaptation

Although it is not unforeseeable that genes related to metal ion binding or transport were under selection and relevant GO terms were over-represented in both species since Mount Etna regularly releases volcanic matter through eruptions, it is interesting that genes related to magnesium, potassium and manganese ions were only selected in the low-elevation *S. chrysanthemifolius*. This could be due to them being highly water-soluble elements, so they would concentrate in lower

elevations as they are washed down by precipitation. It was found that manganese segregate on the surface instead of in the bulk of lava stones collected from Mount Etna and that manganite and manganese oxides are water soluble (Cristaldi, Fortuna, & Gulino, 2013). Leaching of potassium, iron and magnesium was also discovered (Cimino & Toscano, 1998) and high levels of magnesium, manganese and iron were found in potable water in the Mount Etna area (Giammanco, Valenza, Pignato, & Giammanco, 1996). Besides, the presence of iron-related (despite high levels of iron being found to leach) genes being selected in both species could be explained by the fact that iron also occur as insoluble hydroxides easily through oxidation (Giammanco *et al.*, 1996), hence the element exists at high concentrations at high elevations on the mountain as well.

It is surprising that genes related to seed dormancy and high elevations (such as response to ozone and UV-B) were only selected in *S. chrysanthemifolius* but not in the high-elevation *S. aethnensis*. However, the former could be due to *S. chrysanthemifolius* being the more opportunistic species (as discussed above), thus would be dormant and only germinate when favourable conditions arise. One possible explanation for the latter is that higher elevations do not necessarily mean higher UV levels. It was observed that it is often cloudy at high elevations (but not at lower elevations), which would lead to lower UV levels (personal communications with collaborators). Another reaons could be that the ancestral species of the two occupied similar habitats to *S. aethnensis*. Hence, when *S. chrysanthemifolius* adapted to lower elevations, mutations to these genes were required to regulate the presumably strong response to high ozone and UV-B levels. One simple way to test this would be growing the two species in high and low UV settings in the greenhouse and look at the differences in their gene expression patterns. While this study reported genes related to reproduction are strongly selected in LG 4, Chapman *et al.* (2016) did not show any quantitative trait loci on LG 4 that were inflorescence-related. This could be due to that the genes related to reproduction identified in this study did not influence inflorescence development, but other aspects of reproduction such as production of flavonoid and aromatic compounds, and anther and fruit dehiscence. Consistent with the findings in *Chapter 3* (Wong *et al.*, 2020), genes and GO terms related to reproduction, metal ion, and stress response were under selection and over-represented respectively in this study. Other than future experiments explained above, other interesting experiments include those involving anther and fruit dehiscence mechanisms, response to substrates enriched with different metal ions, amount of

chemical compounds produced, and even germination success of each species with or without a dormant period.

4.5.6 Insights into selection and adaptation at high and low elevations

While *Chapter 3* investigated divergent selection in the Etnean *Senecio* system, this chapter looked into positive selection in general in Etnean *Senecio*. Genomic regions under positive selection in both species would likely indicate the presence of divergent selection. As only 76 significant outliers under divergent selection were identified in *Chapter 3* while a lot more regions with sweep signals were identified in this study, it is likely that each species has experienced unique selective forces that resulted in sweeps in different alleles during adaptation to their respective environments.

A few studies have explored positive selection in plants that are adapted to high and low elevations, and discovered candidates that are similar to those found in this study. Regarding defence responses, Buckley *et al.* (2019) found high proportion of indole glucosinolates in high-elevation *Arabidopsis halleri,* which was associated with reduced oviposition from butterflies. However, they pointed out that herbivory could also be a selective agent at low elevations, as low-elevation *A. thaliana* had been found to have reduced production of these chemicals to avoid oviposition from other butterfly species (de Vos, Kriksunov, & Jander, 2008). This could explain why sweep signals were present in genes and GO terms were over-represented related to defence responses in both *S. aethnensis* and *S. chrysanthemifolius*.

Regarding floral development, it was found in six alpine plant species that high elevation populations were more plastic in response to pollinators in terms of flower longevity (suspected to compensate for low pollination at higher elevations), and that they had a higher capacity to senescence quickly after successful pollination to reduce floral costs (Trunschke & Stöcklin, 2017). Studying another six species, Rasmann, Pellissier, Defossez, Jactel, & Kunstler (2014) also showed that production of flavonoid compounds tended to increase at high elevations, potentially related to protection against UV, freezing tolerance and pigmentation. A lot of overlap in biological functions of genes under selection was also observed between this study and those in high-elevation *Arabidopsis*

halleri ssp. gemmifera: enriched G.O. terms in the latter include seed dormancy, photosynthesis, regulation of flower development, transition of meristem from vegetative to reproductive phase, water deprivation, and so on (Kubota *et al.*, 2015). However, some of these were found to be selected in the low-elevation *S. chrysanthemifolius* instead of the high-elevation *S. aethnensis* in this study, such as those related to seed dormancy, response to water deprivation and UV. As discussed above, it is possible that the ancestral species of *S. aethnensis* and *S. chrysanthemifolius* was adapted to habitats similar to the current habitat of *S. aethnensis* and had selective sweeps on genes related to high-elevation adaptations. Hence, new sweeps were required to remove those high-elevation alleles that were not suitable for *S. chrysanthemifolius*; whereas *S. aethnensis* retained certain levels of polymorphisms in those adaptation genes thus no sweeps were detected. A more comprehensive picture of what genes are being selected in which species could be created if a fully annotated genome is available to carry out gene enrichment analyses (in this study, likely not all sweeps signals were true positives, not all genes in regions under sweeps were identified, not all annotations were found for genes identified, and the identified annotations came from a variety of species and families). Hence, future directions should focus on generating a better annotated genome assembly, which would be useful in other genomics studies in the system as well.

In conclusion, this study showed a general positive correlation between recombination rate and nucleotide diversity, fitting theoretical expectations (Begun & Aquadro, 1992). As the first study to carry out a genome-wide scan for positive selection signals in the Etnean *Senecio* system, the current study revealed many regions with putative sweeps spread within and across all linkage groups, with various strengths and ages. A great number of candidate genes were also identified to be under selection at high and low elevations, which could be used in future studies on environmental (extrinsic) selection, a lesser-studied aspect in the system. Building on similar studies of the same kind, this study also provided insights into a previously unstudied selective agent – a volcano which has tremendous influence on environmental conditions and the composition of fauna. In the future, more research can be done regarding plant-environment and plant-fauna interactions to examine the selective pressures in detail. More genomic sequencing projects to re-sequence candidate regions under positive selection, improve the genome assembly and increase sample sizes would also increase the resolution

of genomics studies and reduce the number of false-positive signals for a more precise picture of genomic and selection patterns in this system.

5. Multiple synthetic crosses confirmed hybrid breakdown in Etnean *Senecio* and significant differences from the natural hybrid, *Senecio squalidus* – insights from phenotypic traits

5.1 Abstract

Synthetic hybrids generated in the greenhouse are useful for studying hybrid variations and genetic incompatibilities. They can also be used to pinpoint evolutionary changes in natural hybrids found in the wild since their speciation and adaptation to their respective habitats. In this study, four F_2 families have been generated from crossing *Senecio aethnensis* and *Senecio chrysanthemifolius*, two of which were derived from a reciprocal F_0 cross to investigate potential nucleocytoplasmic incompatibilities. Various previously reported and newly identified signs of hybrid breakdown, such as the presence of leaf trichomes and lack of ray flowers, were observed, pointing to a certainty in incompatibilities between the two species. Comparing various growth, leaf and floral traits, I found that traits were significantly different among individuals of different F_2 families, and within each family there was great variation depending on the hybrid lineage. Besides, significant differences in trait measurements between offspring of different cytoplasmic background have been reported for the first time and suggests the presence of nucleocytoplasmic incompatibilities. Apart from synthetic hybrids, wild leaf samples from the natural hybrid species in the UK, *Senecio squalidus,* were collected. Leaf traits from these samples were analysed with those from synthetic hybrids with the aim to identify any differences that might point to morphological novelties acquired by the natural hybrid species after speciation. Indeed, natural and synthetic hybrids differed in a number of traits but further research is required to identify the genetic basis of the differences. Unlike their Mount Etna progenitors, both synthetic and natural hybrids had very variable leaf shape. Future developmental and genetic studies are necessary to understand the mechanism that determines leaf shapes.

5.2 Introduction

Senecio squalidus is an excellent example of speciation *ex situ*. It was generated by plant materials brought back from Mount Etna to the UK. Its widely accepted putative parents, *S. aethnensis* and *S. chrysanthemifolius*, grow at high- and low-elevations on Mount Etna. Although *S. squalidus* thrives as a species in the UK, its close counterparts, natural hybrids on Mount Etna, were suggested to have low fitness due to strong selection against them. It has already been shown in *Chapter 3* (Wong *et al.*, 2020) that one main factor causing this low reproductivity is strong cumulative selection against hybrids – both extrinsic and intrinsic.

Extrinsic selection in the system has been examined through transplant experiments at different elevations on Mount Etna using *S. aethnensis* and *S. chrysanthemifolius* (Walter *et al.*, 2020b). Walter *et al.* (2020b) observed lower survival and photosynthetic activity after summer in *S. aethnensis* in elevations far from its natural range; while *S. chrysanthemifolius* only performed worse after winter at the transplant site furthest from its range. They also found that overall, the two species differed in both the direction and level of plastic response after being transplanted to other elevations, suggesting that they have each specialised to the habitats in their own range. On the other hand, intrinsic selection (genetic incompatibility) between *S. aethnensis* and *S. chrysanthemifolius* has also been identified previously. Brennan *et al.* (2014) found significant transmission ratio distortion (TRD) in around 27% of loci studied and it likely limited introgression between the two species. They suggested the TRD observed could be caused by various reasons such as meiotic drive. Chapman *et al.* (2016) mapped hybrid breakdown to a few small and one large quantitate trait locus (QTL), and also discovered significant TRD in three genomic regions. Using crosses between *S. squalidus* and its parental species, Brennan *et al.* (2019) observed many loci underlying genomic incompatibilities throughout the genome and that previously identified incompatibility loci between the parental species co-located with some of these loci too.

Selection against hybrids can be manifested in a number of ways, such as failure to germinate or reproduction and growth defects. However, these are hard to detect in the field setting as only plants that survive would be found. Other than estimating the amount of selection, another way to study selection against hybrids is through studying hybrid breakdown in the greenhouse setting, using

synthetic hybrids. In a controlled environment, any breakdown detected would be attributed to intrinsic selection and incompatibilities (selection in the field would likely be higher than that observed in the greenhouse since there would also be extrinsic selection). Many studies have investigated synthetic hybrids between closely-related species to look for hybrid breakdown. For example, low pollen fertility (Martin & Willis, 2010) and chlorosis leading to lethality (Zuellig & Sweigart, 2018) have been found in hybrids between *Mimulus guttatus* and *M. nasutus*; increased susceptibility to herbivores has been suggested in hybrids between *Eucalyptus amygdalina* and *E. risdonii* (Whitham, Morrow, & Potts, 1994). Hybrid breakdown can also occur between intraspecific populations: chlorosis, defected embryo development and others have been identified in hybrids between different *Arabidopsis thaliana* accessions (Plötner *et al.*, 2017); poor growth was found in hybrids between *Oryza sativa* (rice) varieties (Matsubara, Ando, Mizubagash, Ito, & Yano, 2007). In Etnean *Senecio,* Hegarty *et al.* (2009) discovered that germination success dropped to 94% in F_1, 73% in F_2 and only 18% in the F_3 generations before bouncing back in F_4 and F_5 generations, although this is likely due to artificial selection that selected hybrids with high survivorship for later generations (Abbott & Brennan, 2014). They also noticed a decline in survivorship in F_1 hybrids since 15.9% of the individuals were albino (chlorosis). Out of the 147 F_2 hybrids Chapman *et al.* (2016) studied, only 78 (52.3%) individuals were healthy and did not show stunted or necrotic growth; whereas in Brennan *et al.* (2014), 100 out of 120 (83.3%) F_2 hybrids were healthy and did not experience early mortality, stunted growth or remained vegetative. Contrary to these studies, Ross (2010) found no signs of hybrid breakdown in the F_2 hybrids she studied.

Not only are synthetic hybrids useful to study the potential phenotypic and genetic incompatibilities expressed in hybrids, but they are also useful in studying hybrid speciation and domestication of crop species through comparisons with natural hybrids. Trying to recreate three ancient hybrid *Helianthus* species, Rosenthal, Rieseberg, & Donovan (2005) found that it was rare for synthetic hybrids to possess the phenotype of the hybrid species; however, the relative abundance of hybrid species phenotypes in synthetic hybrids also depended on the specific lineages. Zhang *et al.* (2016) created synthetic hybrids using the progenitor parent species of *Brassica napus* (rapeseed) and compare the two's gene expression patterns in order to reveal the changes from hybridisation to later domestication. Using synthetic hybrids, some outstanding questions about hybrid speciation could be

answered, such as how much of the ecological, morphological and genetic novelty observed in the hybrid species actually came from its progenitors through hybridisation, or whether it is due to divergent evolution in their new habitats after speciation (Rieseberg, 1997).

Brennan *et al.* (2014) also uncovered five transmission ratio distortion (TRD) loci that exhibited asymmetric differences in reciprocal hybrids, suggesting potential nucleocytoplasmic incompatibilities, meiotic drive, and/ or gametophytic selection. Any differences in properties in reciprocal hybrids are likely evidence for nucleocytoplasmic interactions, as they have similar nuclear genetic background but different cytoplasmic background (cytoplasm is inherited from one parent only) (Levin, 2003). Reviewing many publications on nucleocytoplasmic interactions in different systems, Levin (2003) found that nucleocytoplasmic conflict could result in a range of variations in reciprocal hybrids, including differences in floral size, colouration, seed yield, vegetative growth rate and pollen fertility. For instance, Leinonen, Remington, & Savolainen (2011) found that *Arabidopsis lyrata* individuals with local cytoplasmic origins were associated with higher fitness in their local habitats; while Zuellig & Sweigart (2018) found that cytoplasmic differences were the basis of chlorosis observed in *Mimulus* hybrids. Levin (2003) also pointed out that even if reciprocal hybrids do not show significant differences in properties, they could still possess nucleocytoplasmic incompatibilities since the conflict could affect both types of hybrids to similar extents.

Studies of *Senecio* hybrid breakdown mentioned above (Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009; Ross, 2010) only used one cross of *S. aethnensis* and *S. chrysanthemifolius* each. To better investigate how widespread hybrid breakdown is phenotypically in the Etnean *Senecio* system, this study reports a large-scale crossing experiment involving more than one F_0 crosses for the first time, to ensure highly comparable results generated by the same person and greenhouse conditions. This is also the first study that attempted to compare natural hybrids of the two species, *S. squalidus*, to synthetic hybrids. Although nucleocytoplasmic incompatibilities have been reported to manifest intrinsically in TRDs (Brennan *et al.*, 2014), they have not been investigated phenotypically. To address the above unexplored areas in this system, I hypothesise that: a) F_2 hybrids would show reduction in fitness, such as low survival rate or stunted growth; b) different genetic background of the F_2 families would result in significant differences in breakdown traits in each family; c) F_2 hybrids which are derived lines of a reciprocal cross would show significant differences in properties due to

nucleocytoplasmic incompatibilities; d) *S. squalidus* would show certain significant differences compared to synthetic F_2 hybrids; e) any differences between field-collected and greenhouse-grown plants of the same genetic background would be due to phenotypic plasticity influenced by environmental conditions. To test the above hypotheses, four F_2 hybrid families were selected to study leaf traits, character differentiation and hybrid breakdown; two of which were selected as they were derived lines of a reciprocal cross to allow tests of nucleocytoplasmic incompatibilities. Traits of natural and synthetic hybrids were also compared to look for changes the natural hybrid has potentially acquired after speciation.

5.3 Materials and Methods

5.3.1 Crossing design

Seeds from pure *S. aethnensis* and *S. chrysanthemifolius* used for generating hybrids were collected by Greg Walter on Mount Etna and subsequently germinated by me in the greenhouse of the Department of Plant Sciences, University of Oxford. Crosses between reproductive individuals of the two species were carried out between winter 2018 to spring 2019. Seeds of successful crosses were germinated and grown starting from summer 2019. Four F_1 crosses with substantial number of seeds produced were then used to establish respective F_2 families, which were germinated and grown between winter 2019 and spring 2020. Details of individuals used in crosses are in Table 5.1.

Table 5.1. Parental species information of F_2 families used in this study. Crosses were carried out by rubbing two mature capitula (enclosed before manual pollination) against each other, then enclosed in mesh bags until seed set. Each F_2 family has known F_0 and F_1 maternal and paternal individuals. Subsequent germination and survival details in F_1 are provided in Table 5.2.

F_2 family	**F_0 parents**			
	♂ species	♂ Individual ID	♀ species	♀ Individual ID
B71	*S. aethnensis*	2103.3	*S. chrysanthemifolius*	2097.3
B87	*S. chrysanthemifolius*	2097.3	*S. aethnensis*	2103.3
193B	*S. chrysanthemifolius*	2100.1	*S. aethnensis*	2104.1
B228	*S. aethnensis*	2106.1	*S. chrysanthemifolius*	2098.3

5.3.2 Seed germination and plant rearing conditions

Seeds were germinated in soil, comprising 3:1 mixture of compost and medium grade vermiculite, in standard 8 x 5 tray pots. Seedlings were transplanted to 12 cm pots once they were established, using soil of the same composition as that for germination. Plants were then grown in controlled conditions with a 16-hour photo period (LED lighting) at 20°C.

To cross individuals, mature capitula were rubbed together, enclosed with cellulose mesh bags then tied. All flower heads were enclosed before artificial fertilisation. In F_0 and F_1 crosses, each bag of crossed capitula only consisted of crossing with those of one other individual thus seeds generated would have known male and female parents. In F_0, 12 and 21 *S. aethnensis* and *S.*

chrysanthemifolius plants were grown, creating 70 F_1 families upon successful fertilisation. 26 of these F_1 families had at least 5 seeds and were subsequently germinated for generating seeds for F_2 families. Two to eight individuals survived to reproductive maturity in each F_1 family. Individuals within the same F_1 family were crossed, generating multiple F_2 families with sufficient seeds (> 50) for germination. Among these, four F_2 families were chosen for this study (Table 5.1 and 5.2). Two of these families were derived lines of a reciprocal F_0 cross. This was to test for effect of nucleocytoplasmic background on traits. In F_2 crosses, each bag of crossed capitula consisted of crossing with those from multiple random individuals within the same family to enhance fertilisation success, thus only the female parent of resulting seeds is known. Etnean *Senecio* are highly self-incompatible (Chapman, Forbes, & Abbott, 2005) thus there was no need to emasculate the capitula before crossing. Hence all crosses were reciprocal as each capitulum was both the donor and recipient of pollen.

All F_0 *S. aethnensis* and *S. aethnensis,* individuals within the same F_1 and F_2 families were grown in the same greenhouse bay. However, two greenhouse bays were used for growing F_1 and F_2 families due to size restrictions.

5.3.3 *Senecio squalidus* samples

To compare leaf traits of natural hybrids and synthetic hybrids described above, *S. squalidus* leaves were collected from four locations in the UK – Oxford, Exeter, Newcastle and Edinburgh. Two to four leaves in good conditions were collected for each individual and photographed on gridded paper for computational analysis. Only undamaged leaves were collected from the main stem of the plants, and different sizes of leaves were collected to represent more or less the average leaf of each plant. All plants used in leaf collection were in their flowering stage (not pre-reproductive or fruiting stages).

To compare leaf traits of plants with identical genetic composition grown in the field and greenhouse, cuttings from samples in Exeter were also collected and grown in the greenhouse in the

Department of Plant Sciences, University of Oxford. Leaves were collected and photographed in the same protocol as above from mature plants.

5.3.4 Leaf traits

At the first sign of flower buds in each F_2 individual, three mid-stem leaves were randomly chosen and photographed on gridded paper. Leaf images were scaled individually and analysed in LeafProcessor (Backhaus *et al.*, 2010). Various leaf morphometric parameters were estimated by the software, including perimeter, area, length, width, compactness (ratio of perimeter squared and area).

Leaves were then quickly snap-frozen in liquid nitrogen and subsequently stored in -80°C for potential genetic studies in the future.

5.3.5 Other phenotypic characterisation in F_2 individuals

Various phenotypic traits were measured manually with the aim to identify any signs of hybrid breakdown in F_2 individuals grown in the greenhouse. Phenotypic characters measured include presence of trichomes on leaves, plant height at first anthesis (cm), number of effective branches (branches with two or more nodes) at first anthesis, stem diameter at first anthesis (mm), number of ray flowers on apical capitulum, number of capitula on the apical umbel, width of receptacle of apical capitulum (disc diameter; mm), and length of pedicel of apical capitulum (mm).

5.3.6 Statistical analyses

Six groups of data were analysed in this study: (a) F_2 individuals and their respective progenitors in F_0 and F_1 generations; (b) all F_2 individuals that comprises the four families; (c) the two F_2 families which are derived lines of a reciprocal F_0 cross; (d) all F_2 individuals and *S. squalidus* samples collected in the field and grown in greenhouse; (e) all F_2 individuals and *S. squalidus* grown in the

greenhouse (from cuttings collected from Exeter); and (f) *S. squalidus* collected from Exeter and their cuttings grown in the greenhouse.

In the leaf measurements, entries with leaf area below 5 or above 400 mm^2 were outliers and treated as measurement error and removed from the dataset (Minimum, maximum, mean and standard deviations of measurements can be found in Table 5.4).

Normality of the datasets were tested using the R package *rstatix* (Kassambara, 2020). Given the datasets did not show normal distribution, a non-parametric test of multivariate analysis of variance was carried out using the R package *npmv* (Ellis, Burchett, Harrar, & Bathke, 2017) on data groups (b) to (f). Summary statistics, including mean, standard deviation, minimum and maximum values, were obtained using the R package *psych* (Revelle, 2019), in each of the F_2 families, all F_2 individuals, all *S. squalidus* individuals, *S. squalidus* Exeter samples collected in the field and *S. squalidus* Exeter samples grown in the greenhouse. Kruskal-Wallis test, Wilcoxon rank-sum tests, and Spearman's correlation tests were carried out using the R package *stats*. Kruskal-Wallis tests were carried out in data groups (b), (d), (e) and (f). Subsequent pairwise Wilcoxon rank-sun tests were carried out in data group (b). Spearman's correlation tests were carried out between all pairs of traits measured in data group (b) and all *S. squalidus* individuals respectively. Principal component analyses (PCA) were carried out using the R package *ggbiplot* (Vu, 2011) in all six data groups. When analyses only involved F_2 individuals (both leaf and other traits), leaf measurements were averaged among the three leaves measured for each individual. In other analyses, all three leaf measurements for each F_0, F_1, F_2, and *S. squalidus* individuals were used.

5.4 Results

To compare different growth, leaf and floral traits between different groups of interests, measurements were taken on the plants collected in the field or grown in the greenhouse. Only leaf traits were measured in *S. squalidus* samples since they were at different stages of flowering at the time of sampling. Leaf traits, alongside growth and floral traits were measured in F_2 individuals generated from the crossing experiment in the greenhouse. Growth and floral traits were not used for the F_0 and F_1 progenitors of F_2 plants as their sample sizes (each F_2 family have only four F_0 and two F_1 progenitors) are too small to allow meaningful comparisons.

5.4.1 Survival and signs of hybrid breakdown

Number of individuals that survived to maturity was variable among the four F_2 families, ranging from 50.85% to 82% (Table 5.2). Only two out of four F_2 families had individuals with early mortality (Figure 5.1) – 1.69% and 22% in families B87 and 193B respectively (Table 5.2). These individuals all grew to maturity and had flower buds before they died. However, they either died before the flower buds bloomed or shortly after they started blooming (before they could set seeds). Some signs of hybrid breakdown were also present in F_2 individuals, including presence of trichomes on leaves and capitula that lack ray florets (Figure 5.1).

Table 5.2. Information on germination and survival in the four F_2 families used in this study. Of the individuals grown to maturity in F_1, two were crossed to create the F_2 family. All crosses were carried out by rubbing two mature capitula (enclosed before manual pollination) against each other, then enclosed in mesh bags until seed set.

F_2 family	Respective F_1 families				F_2 individuals				
	F_1 cross code	No. of seeds sowed	No. germinated and grown to maturity	% grown to maturity	No. of seeds sowed	No. grown to maturity	% grown to maturity	No. of early mortalities	% early mortality
B71	AC02	10	6	60	64	39	60.94	0	0
B87	AC03	10	6	60	59	30	50.85	1	1.69
193B	AC15	10	4	40	50	41	82	11	22
B228	AC20	10	6	60	64	40	62.5	0	0

(a)



(b)



(c)



(d)



Figure 5.1. Photos showing trait variabilities and signs of hybrid breakdown. (a) Leaf shape variation in the same F_2 family; (b) left: leaf without trichomes (parental phenotype); right: F_2 with trichomes; (c) A F_2 plant showing early mortality; (d) F_2 capitula that lack ray florets.

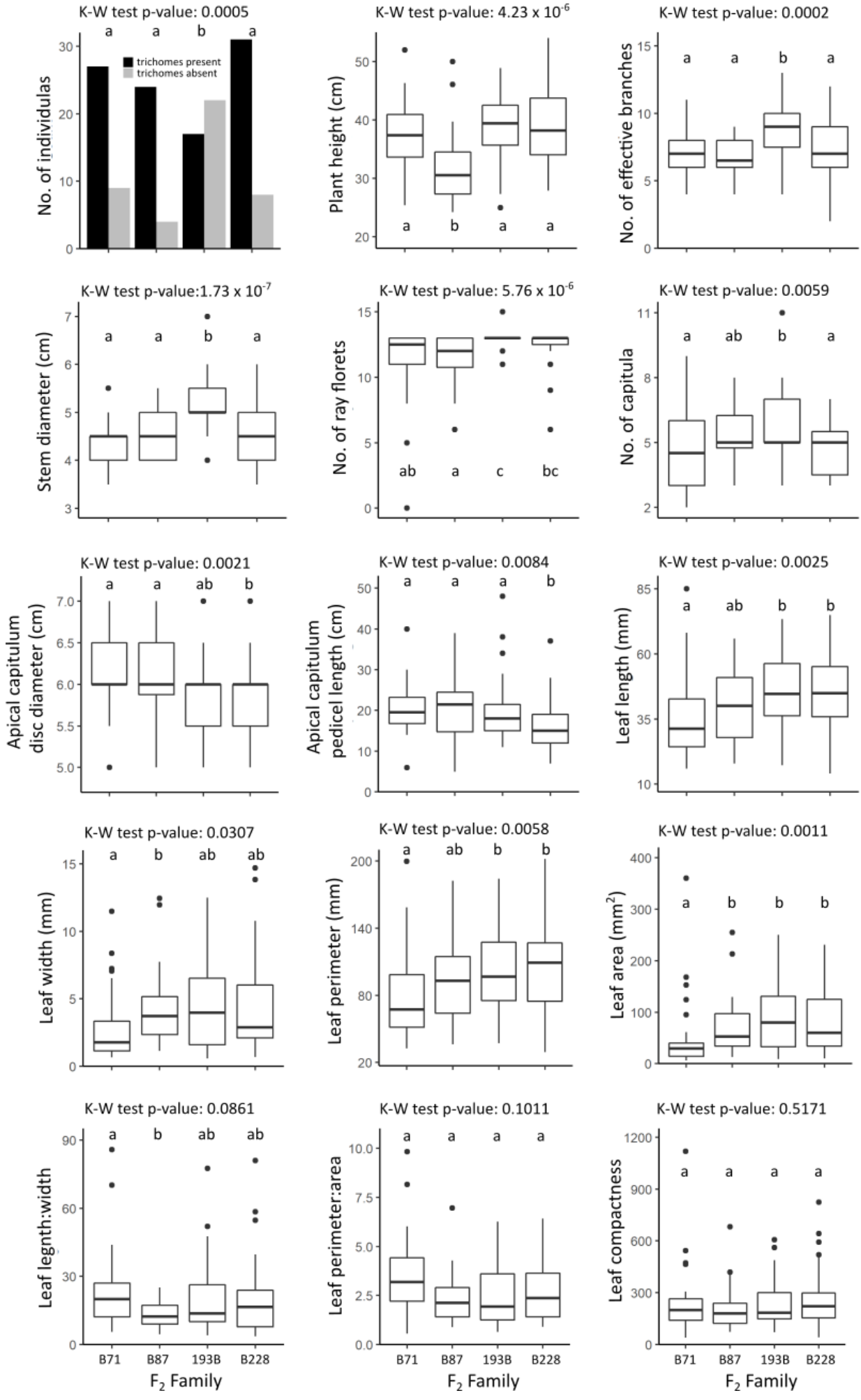

K-W test p-value: 0.0005
a a b a
trichomes present
trichomes absent
No. of individulas
K-W test p-value: 4.23 x 10-6
a b a a
Plant height (cm)
K-W test p-value: 0.0002
a a b a
No. of effective branches
K-W test p-value:1.73 x 10-7
a a b a
Stem diameter (cm)
K-W test p-value: 5.76 x 10-6
ab a c bc
No. of ray florets
K-W test p-value: 0.0059
a ab b a
No. of capitula
K-W test p-value: 0.0021
a a ab b
Apical capitulum disc diameter (cm)
K-W test p-value: 0.0084
a a a b
Apical capitulum pedicel length (cm)
K-W test p-value: 0.0025
a ab b b
Leaf length (mm)
K-W test p-value: 0.0307
a b ab ab
Leaf width (mm)
K-W test p-value: 0.0058
a ab b b
Leaf perimeter (mm)
K-W test p-value: 0.0011
a b b b
Leaf area (mm2)
K-W test p-value: 0.0861
a b ab ab
Leaf legnth:width
K-W test p-value: 0.1011
a a a a
Leaf perimeter:area
K-W test p-value: 0.5171
a a a a
Leaf compactness
B71 B87 193B B228
F2 Family

Figure 5.2. Box plots of each phenotypic trait quantified in the four F_2 families. P-values for Kruskal-Wallis test are indicated at the top of each plot; letters above each box represent groupings following results of Wilcoxon rank-sum test. Detailed results of the two tests can be found in the supplementary tables.

5.4.2 Differences in phenotypic characteristics among F_2 groups

Non-parametric multivariate analysis of variance (MANOVA) revealed that traits are significantly different among the four F_2 families (p-value < 0.0001; Table 5.3). Summary statistics of each trait for each F_2 family can be found in Table 5.4. Kruskal-Wallis tests revealed that all traits, except leaf length to width ratio, leaf perimeter to area ratio and leaf compactness, are significantly different among the four F_2 families (Table 5.5). Pairwise rank-sum tests (Table 5.6) and box plots (Figure 5.2) further revealed family pairs that had significantly different trait measurements. For instance, family 193B had significantly fewer individuals with leaf trichomes, more effective branches (branches with two or more nodes) and larger stem diameter than the other three families; family B71 had significantly smaller leaf area than the other three families; family B87 individuals were significantly shorter than the other three families.

Table 5.3. Results of non-parametric multivariate analysis of variance (MANOVA).

Type of test	Test statistic	df1	df2	P-value
Among all F_2 families	5.55	18.90	848.67	< 0.0001
Between F_2 families B71 and B87 only	4.62	6.90	415.15	< 0.0001
Between F_2 hybrids and *S. squalidus*	154.64	2.24	2208.83	< 0.0001
Between F_2 hybrids and *S. squalidus* – Exeter samples grown in the greenhouse	47.50	2.39	886.89	< 0.0001
Between *S. squalidus* - Exeter samples collected in the field and grown in the greenhouse	8.88	2.56	747.33	< 0.0001

Spearman's correlation analysis (Table 5.7) did not show a clear correlation pattern between traits. Among the floral traits (number of ray florets, number of capitula, apical capitulum disc diameter, apical capitulum pedicel length), only the number of capitula was negatively correlated with apical capitulum pedicel length. All other comparisons were not significantly correlated. Among the leaf traits (length, width, perimeter, area, length to width ratio, perimeter to area ratio, compactness),

all pairs were significantly correlated positively or negatively, except for compactness with length, width and perimeter respectively. Among other vegetative traits (plant height, number of effectively branches and stem diameter), all pairs were significantly and positively correlated except for plant height and number of effective branches. Some size traits were also significantly but negatively correlated each other, such as stem diameter and apical capitulum disc diameter.

PCA using all F_2 individuals and their F_1 and F_0 progenitors showed that F_1 and F_2 hybrids are clearly located in between points representing the two species of F_0 (Figure 5.3, loading values in Table 5.8). However, PCA also showed that each cluster of F_2 family were not clearly distinct (Figure 5.4; loading values in Table 5.8).

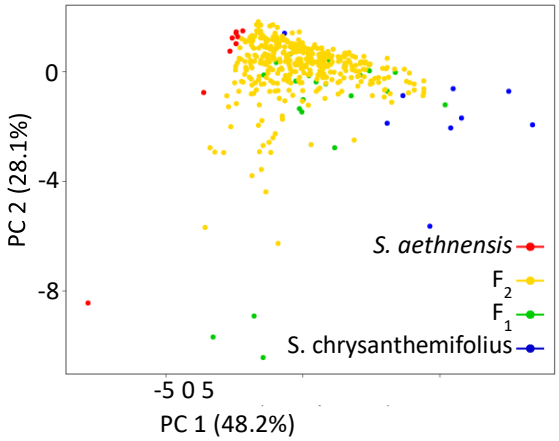


Figure 5.3. PCA of all F_2 individuals used in this study, and their respective parents in the previous two generations.

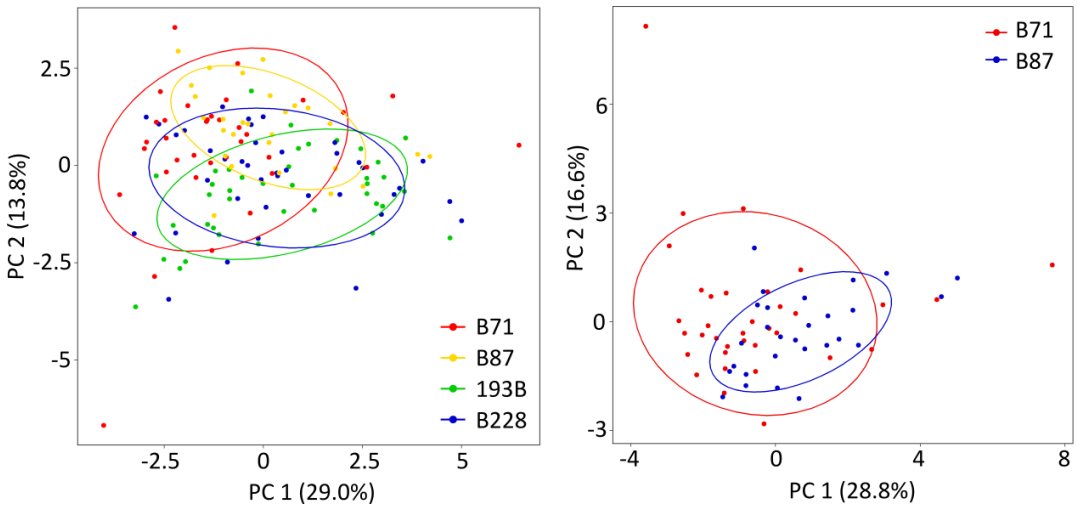


Figure 5.4. PCA of (left) the four F_2 families and (right) the two F2 families which have the same F_0 parents but opposite crossing direction (different nucleocytoplasmic background.

Table 5.4. Summary statistics for the traits measured in each group.

Trait	B71 (N = 39)				B87 (N = 29)				193B (N = 39)				B228 (N = 40)			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Plant height (cm)	37.52	6.04	25.4	52	31.93	6	24.2	50	39.25	5.34	25	48.9	38.85	5.98	27.9	54
No. effective branches	7.13	1.65	4	11	6.72	1.6	4	10	8.74	2.15	4	13	7.18	2.22	2	12
Stem diameter (cm)	4.41	0.43	3.5	5.5	4.62	0.55	4	5.5	5.18	0.54	4	7	4.64	0.62	3.5	6
No. ray florets	11.46	2.7	0	13	11.28	2.05	6	13	12.95	0.51	11	15	12.45	1.34	6	13
No. capitula	4.63	1.6	2	9	5.45	1.43	3	8	5.79	1.76	3	11	4.69	1.34	3	7
Apical capitulum disc diameter (cm)	6.16	0.5	5	7	6.09	0.49	5	7	5.88	0.48	5	7	5.73	0.51	5	7
Apical capitulum pedicel length (cm)	19.64	6.53	6	40	20.89	7.76	5	39	19.74	7.6	11	48	16.21	6.07	7	37
Leaf area (mm^2)	54.29	69.6	6.17	360.4	71.58	56.03	12.84	255	91.7	68.81	8.88	250.3	83.99	65.11	10.06	231.4
Leaf length (mm)	37.16	17.07	15.98	85.01	40.54	13.44	17.87	65.85	45.44	13.4	17.3	73.35	46.11	15.09	14.1	74.96
Leaf perimeter (mm)	82.3	40.19	32.65	199.7	95.01	36.16	36.17	182.4	101.9	35.41	37.18	184.1	107	43.32	29.38	201.8
Leaf width (mm)	3.05	2.61	0.67	11.48	4.57	3.07	1.14	12.44	4.47	3.38	0.58	12.49	4.35	3.47	0.68	14.7
Leaf length:width	22.76	16.63	5.65	85.73	13.11	5.72	3.94	25.09	21.04	16.09	4.05	77.51	19.97	16.23	3.71	80.91
Leaf perimeter:area	3.39	1.96	0.55	9.82	2.6	1.64	0.88	7.72	2.56	1.68	0.65	6.26	2.74	1.6	0.9	6.41
Leaf compactness	251.6	184.7	40.42	1119	242.6	205.4	73.36	1065	229.4	129.4	71.03	607.5	266.2	169.4	42.05	824.3

Table 5.4. continued.

Trait	All F_2 individuals (N = 379)				All *S. squalidus* samples (N = 707)				*S. squalidus* Exeter/ field (N = 182)				*S. squalidus* Exeter/ greenhouse (N = 124)			
	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Leaf area (mm^2)	77.9	81.61	6.07	387.3	65.71	69.48	5.05	368.2	42.77	47.47	5.29	265.8	56.64	61.71	5.2	368.2
Leaf length (mm)	43.35	15.28	12.82	87.02	27.54	14.19	4.86	81.36	20.11	9.78	4.89	62.53	26.91	13.05	5.21	68.27
Leaf perimeter (mm)	99.03	40.43	26.92	230	71.69	44.63	11.7	295.1	53.76	34.29	13.96	211.9	67.22	37.54	15.33	220.6
Lleaf width (mm)	4.2	3.59	0.28	20.57	5.94	4.7	0.4	29.63	5.19	4.49	0.46	29.63	5.05	4.03	0.41	23.6
Leaf length:width	19.15	20.08	2	183.9	7.3	7.42	0.95	70.01	6.22	6.6	0.95	42.89	9.36	10.28	0.97	70.01
Leaf perimeter:area	2.79	2.73	0.33	21.08	2.14	2.4	0.31	22.2	2.16	1.76	0.44	11.38	2.28	2.19	0.37	16.32
Leaf compactness	253.2	289.6	30.29	2931	158.1	291.5	15.98	3031	118.6	166.8	15.98	1695	152.8	235.7	21.61	2251

Table 5.5. Results of the Kruskal-Wallis test. *indicates significant P-values.

	Among four F_2 families (df = 3)		All F_2 & *S. squalidus* (df = 1)		All F_2 & *S. squalidus* Exeter/ greenhouse (df = 1)		*S. squalidus* Exeter/ field & greenhouse (df = 1)	
Trait	**χ^2**	**P - value**	**χ^2**	**P - value**	**χ^2**	**P - value**	**χ^2**	**P - value**
Presence of trichomes	17.91	0.00046*						
Plant height	27.67	4.25×10^{-06}*						
No. effective branches	19.45	0.00022*						
Stem diameter	34.28	1.73×10^{-07}*						
No. ray florets	27.04	5.76×10^{-06}*						
No. capitula	12.49	0.0059*						
Apical capitulum disc diameter	14.69	0.0021*						
Apical capitulum pedicel length	11.71	0.0084*						
Leaf length	14.36	0.0025*	234.42	$< 2.2 \times 10^{-16}$*	96.88	$< 2.2 \times 10^{-16}$*	22.92	1.69×10^{-06}*
Leaf width	8.90	0.031*	48.03	4.20×10^{-12}*	5.73	0.017*	0.032	0.86
Leaf perimeter	12.52	0.0058*	126.85	$< 2.2 \times 10^{-16}$*	60.67	6.76×10^{-15}*	13.80	0.00020*
Leaf area	16.15	0.0011*	6.03	0.014*	7.46	0.0063*	3.76	0.052
Leaf length:width	6.59	0.086	310.19	$< 2.2 \times 10^{-16}$*	76.39	$< 2.2 \times 10^{-16}$*	13.94	0.00019*
Leaf perimeter:area	6.23	0.10	22.11	2.58×10^{-06}*	2.53	0.11	0.00011	0.99
Leaf compactness	2.28	0.52	161.2	$< 2.2 \times 10^{-16}$*	48.94	2.65×10^{-12}*	7.01	0.0081*

Table 5.6. P-values for pairwise Wilcoxon rank-sum tests, Bonferroni-corrected, among the four F_2 families. *indicates significant P-values.

Presence of trichomes			
	B71	B87	193B
B87	1	-	-
193B	0.0486*	0.0033*	-
B228	1	1	0.0074*

Plant height			
	B71	B87	193B
B87	0.0015*	-	-
193B	0.8278	2.20×10^{-05}*	-
B228	1	5.30×10^{-05}*	1

No. of effective branches			
	B71	B87	193B
B87	1	-	-
193B	0.00619*	0.00032*	-
B228	1	1	0.02256*

Stem diameter			
	B71	B87	193B
B87	0.6599	-	-
193B	2.50×10^{-08}*	0.001*	-
B228	0.3003	1	0.0034*

No. of ray florets			
	B71	B87	193B
B87	1	-	-
193B	0.00013*	3.70×10^{-05}*	-
B228	0.10361	0.03327*	0.12789

No. of capitula			
	B71	B87	193B
B87	0.218	-	-
193B	0.062	1	-
B228	1	0.135	0.03*

Apical capitulum disc diameter			
	B71	B87	193B
B87	1	-	-
193B	0.127	0.5798	-
B228	0.0047*	0.0407*	1

Apical capitulum pedicel length			
	B71	B87	193B
B87	1	-	-
193B	1	1	-
B228	0.032*	0.039*	0.091

Leaf length			
	B71	B87	193B
B87	0.4663	-	-
193B	0.0045*	1	-
B228	0.0062*	1	1

Leaf width			
	B71	B87	193B
B87	0.019*	-	-
193B	0.2	1	-
B228	0.179	1	1

Leaf perimeter			
	B71	B87	193B
B87	0.341	-	-
193B	0.011*	1	-
B228	0.01*	1	1

Leaf area			
	B71	B87	193B
B87	0.0197*	-	-
193B	0.0051*	1	-
B228	0.0033*	1	1

Leaf length:width			
	B71	B87	193B
B87	0.025*	-	-
193B	1	0.857	-
B228	1	0.99	1

Leaf perimeter:area			
	B71	B87	193B
B87	0.11	-	-
193B	0.22	1	-
B228	1	1	1

Leaf compactness			
	B71	B87	193B
B87	1	-	-
193B	1	1	-
B228	1	0.72	1

Table 5.7. Spearman's correlation between all pairs of traits measured in all F_2 individuals. Values below the diagonal are Spearman's Rho; values above are the P-values.

Presence. trichomes	0.1106	0.8712	0.0386	0.0706	0.1511	0.8544	0.8096	0.1425	0.0490	0.0724	0.1184	0.0274	0.0689	0.3068
-0.13	**Plant height**	0.1775	6.26E-06	0.0086	0.8585	0.0550	0.2505	0.0647	0.5987	0.1949	0.4345	0.1994	0.5556	0.7030
-0.01	0.11	**No. effective branches**	0.0019	0.4280	0.8918	0.0017	0.1644	0.1980	0.5355	0.2425	0.5230	0.1980	0.6457	0.7984
-0.17	0.37	0.26	**Stem diameter**	0.0053	0.0618	0.0127	0.2292	0.0044	0.0263	0.0174	0.0059	0.2869	0.0406	0.7579
-0.15	0.22	0.07	0.23	**No. ray florets**	0.9311	0.8460	0.0582	0.0994	0.9290	0.1902	0.4461	0.2729	0.7568	0.2480
-0.12	-0.02	-0.01	0.16	-0.01	**No. capitula**	0.6361	0.0019	0.0443	0.2805	0.0698	0.0369	0.6804	0.0394	0.3070
0.02	-0.16	-0.26	-0.21	-0.02	0.04	**Capitulum disc diam.**	0.2520	0.5004	0.3628	0.4907	0.2618	0.5004	0.2023	0.5263
-0.02	-0.10	0.12	-0.10	-0.16	-0.26	0.10	**Capitulum pedicel length**	0.0009	0.0966	0.0020	0.0042	0.0009	0.0242	0.7415
0.12	0.16	0.11	0.24	0.14	0.17	-0.06	-0.28	**Leaf length**	$<2.2x10^{-16}$	$<2.2x10^{-16}$	$<2.2x10^{-16}$	0.0002	$2.04x10^{-09}$	0.0676
0.17	-0.04	0.05	0.19	-0.01	0.09	-0.08	-0.14	0.68	**Leaf width**	$<2.2x10^{-16}$	$<2.2x10^{-16}$	$<2.2x10^{-16}$	$<2.2x\ 10^{-16}$	0.0563
0.15	0.11	0.10	0.20	0.11	0.15	-0.06	-0.26	0.98	0.77	**Leaf perimeter**	$<2.2x10^{-16}$	$<2.2x10^{-16}$	$8.51x10^{-10}$	0.0584
0.13	0.07	0.05	0.23	0.06	0.18	-0.10	-0.24	0.82	0.88	0.84	**Leaf area**	$<2.2x10^{-16}$	$<2.2x10^{-16}$	0.002571
-0.19	0.11	0.11	-0.09	0.09	-0.03	-0.06	-0.28	-0.31	-0.85	0.98	-0.63	**Leaf length:width**	$<2.2x10^{-16}$	$9.37x10^{-06}$
-0.15	-0.05	-0.04	-0.17	-0.03	-0.17	0.11	0.19	-0.48	-0.65	-0.49	-0.76	0.60	**Leaf perimeter: area**	$<2.2x10^{-16}$
-0.09	0.03	0.02	-0.03	0.10	-0.09	0.05	0.03	0.15	-0.16	0.16	-0.25	0.37	0.74	**Leaf compactness**

5.4.3 Differences in phenotypic characteristics between the two F_2 families derived from a reciprocal F_0 cross

Non-parametric MANOVA revealed that traits are significantly different between F_2 families B71 and B87 (p-value < 0.0001; Table 5.3). Pairwise rank-sum tests (Table 5.6) and box plots (Figure 5.2) further revealed that these two families only had four significantly different trait measurements, among those none were floral traits. Family B71 was significantly taller, had significantly shorter leaf width, smaller area, and higher leaf length to width ratio. PCA also did not show clear clusters of the two family (Figure 5.5; loading values in Table 5.8).

5.4.4 Differences in leaf traits between synthetic F_2 hybrids and the hybrid species *S. squalidus*

A total of 707 *S. squalidus* leaves were analysed. All leaf traits were significantly correlated with each other (Table 5.9) These traits were compared with the F_2 hybrids above.

Non-parametric MANOVA showed that the two groups are significantly different (p-value < 0.0001; Table 5.3). Statistics of leaf traits of all *S. squalidus* and F_2 individuals are summarised in Table 5.4. Kruskal-Wallis test revealed that *S. squalidus* and F_2 hybrids were significantly different in all the traits (Table 5.5). *S. squalidus* had significantly shorter leaves, wider leaves, shorter perimeter, smaller area, smaller length to width ratio, smaller perimeter to area ratio and lower compactness than F_2 hybrids (Figure 5.5). Although *S. squalidus* and F_2 hybrids were significantly different overall, PCA did not show clear clustering of the two groups. Many F_2 individuals seem to be segregating towards *S. squalidus* in the PCA, although some *S. squalidus* occupied areas not overlapped with any F_2 individuals (Figure 5.6; loading values in Table 5.8).

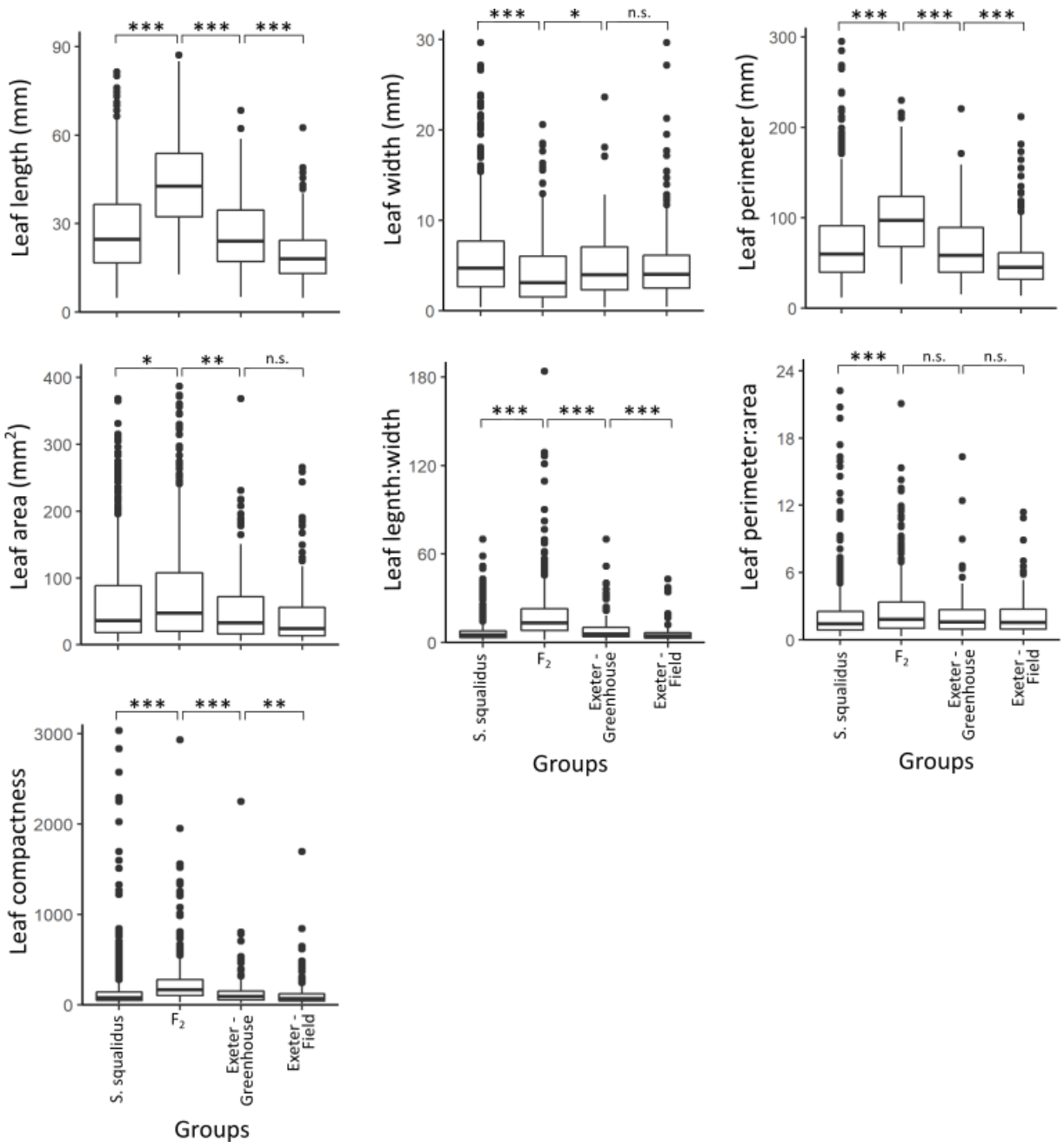


Figure 5.5. Box plots of each leaf phenotypic trait quantified in the field collected *S. squalidus*, F_2 individuals, *S. squalidus* – Exeter samples grown in the greenhouse and collected in the field. * indicates significance of each pairwise Kruskal-Wallis test (*: p-value < 0.05; **: p-value < 0.005; ***: p-value < 0.0005; n.s. = not significant). Detailed results of the test can be found in the supplementary tables.

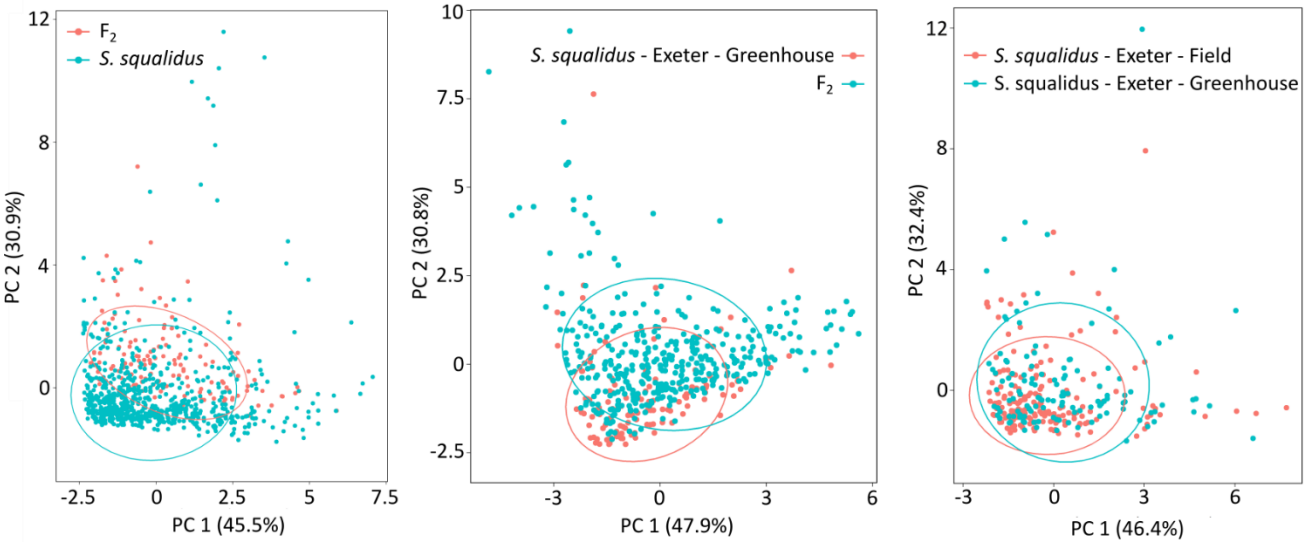


Figure 5.6. PCAs involving F_2 individuals and the hybrid species *S. squalidus*. (left) all F_2 and *S. squalidus* individuals collected from the field and greenhouse; (middle) all F_2 and *S. squalidus* individuals grown in the greenhouse; (right) *S. squalidus* individuals collected in Exeter and later grown in the greenhouse from cuttings.

Table 5.8. Loading values of each principal component in different PCAs.

Trait	Among F_0, F_1 and F_2		Among four F_2 families		F_2 families B71 & B87		All F_2 and *S. squalidus*		All F_2 and *S. squalidus* Exeter/ greenhouse)		*S. squalidus* Exeter/ field & greenhouse	
	PC 1	**PC 2**	**PC1**	**PC2**	**PC1**	**PC2**	**PC 1**	**PC 2**	**PC1**	**PC2**	**PC1**	**PC2**
Presence of trichomes			0.09	0.15	0.14	0.04						
Plant height			0.05	-0.31	-0.06	0.13						
No. effective branches			0.04	-0.27	0.02	0.36						
Stem diameter			0.13	-0.33	0.09	0.20						
No. ray florets			-0.04	-0.22	-0.16	0.01						
No. capitula			0.10	0.04	0.11	-0.13						
Apical capitulum disc diameter			-0.09	0.12	-0.08	0.02						
Apical capitulum pedicel length			-0.15	0.02	-0.16	0.00						
Leaf area	0.47	0.02	0.44	-0.04	0.45	0.09	0.50	0.16	0.49	0.01	0.48	-0.20
Leaf length	0.43	-0.27	0.39	-0.29	0.38	0.32	0.47	-0.23	0.39	0.38	0.46	0.23
Leaf perimeter	0.48	-0.30	0.42	-0.25	0.41	0.29	0.52	-0.22	0.44	0.37	0.53	0.17
Leaf width	0.44	-0.14	0.42	-0.02	0.42	0.06	0.46	-0.01	0.44	0.08	0.48	-0.08
Leaf length:width	-0.29	-0.30	-0.30	-0.37	-0.26	0.42	-0.17	-0.36	-0.27	0.34	-0.16	0.38
Leaf perimeter:area	-0.28	-0.56	-0.36	-0.32	-0.34	0.35	-0.17	-0.60	-0.33	0.49	-0.12	0.61
Leaf compactness	-0.10	-0.64	-0.15	-0.50	-0.16	0.54	0.02	-0.62	-0.17	0.60	0.12	0.59

Table 5.9. Spearman's correlation between all pairs of traits measured in all *S. squalidus* individuals. Values below the diagonal are Spearman's Rho; values above are the p-values.

Leaf length	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	0.0007	7.25×10^{-15}	$<2.2 \times 10^{-16}$
0.60	**Leaf width**	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	$<2.2 \times 10^{-16}$	0.0030
0.95	0.76	**Leaf perimeter**	$<2.2 \times 10^{-16}$	0.0054	4.75×10^{-13}	$<2.2 \times 10^{-16}$
0.70	0.72	0.71	**Leaf area**	1.46×10^{-12}	$<2.2 \times 10^{-16}$	2.37×10^{-07}
0.13	-0.69	-0.10	-0.26	**Leaf length:width**	$<2.2 \times 10^{-16}$	3.34×10^{-13}
-0.29	-0.47	-0.27	-0.84	0.32	**Leaf perimeter:area**	$<2.2 \times 10^{-16}$
0.45	0.11	0.50	-0.19	0.27	0.66	**Leaf compactness**

5.4.5 Differences in leaf traits between synthetic F_2 hybrids and *S. squalidus* - Exeter samples grown in the greenhouse

A total of 124 greenhouse-grown *S. squalidus* leaves were analysed. Non-parametric MANOVA showed that they are significantly different from F_2 individuals grown in the greenhouse (p-value < 0.0001; Table 5.3). Statistics of leaf traits of these two groups are summarised in Table 5.4. Kruskal-Wallis tests showed that the two groups were significantly different in all traits except leaf perimeter to area ratio (Table 5.5). F_2 individuals had significantly longer leaves, narrower leaves, longer leaf perimeter, larger area, larger length to width ratio and higher compactness (Figure 5.3). PCA did not show clear clustering of the two groups, but *S. squalidus* samples occupied some areas not overlapped with F_2 individuals (Figure 5.6; loading values in Table 5.8).

5.4.6 Differences in leaf traits between wild and greenhouse-grown *S. squalidus* under identical genetic background

There were 182 and 124 leaves of *S. squalidus* Exeter samples collected in the field and grown in the greenhouse respectively. Non-parametric MANOVA showed that they are significantly different from each other (p-value < 0.0001; Table 5.3). Statistics of leaf traits of both groups are summarised in Table 5.4. Kruskal-Wallis tests revealed that the two groups are only significantly different in four traits (Table 5.5): the greenhouse grown leaves had significantly longer leaves, longer perimeter, larger length to width ratio and higher compactness (Figure 5.3); while leaf width, area and perimeter to area ratio were not significantly different.

5.5 Discussion

In this study, I set up a crossing experiment with the aim to quantify the extent of hybrid breakdown in four families of F_2 hybrids between *S. aethnensis* and *S. chrysanthemifolius.* I have also measured various traits related to growth, leaves and flowers. I examined these among the four families, and also between F_2 plants and the hybrid species *S. squalidus*. Comparing traits of different groups, I investigated the effects of nucleocytoplasmic and environmental differences on phenotypes, and inferred the consequences of hybrid breakdown on the natural hybrid zone on Mount Etna.

5.5.1 Signs of hybrid breakdown

Synthetic hybrids between *S. aethnensis* and *S. chrysanthemifolius* showed various signs of hybrid breakdown, both previously reported and new. As a direct measure of fitness, germination and survival of hybrids in this study is lower than previously reported: 40 - 60% in F_1 and as low as 50.85% in F_2 hybrids are reported in this study, compared to 94% and 73% in F_1 and F_2 respectively in Hegarty *et al.* (2009). Low fitness in early generation hybrids is expected, as they harbour the most unfavourable association of alleles compared to later generation ones (Arnold & Bennett, 1993; Fishman & Willis, 2001; Johansen-Morris & Latta, 2006). Although no hybrids in this study were albino, had stunted growth or remained vegetative, early mortality was observed in two F_2 families. New signs of hybrid breakdown, including lack of ray florets on capitula and presence of leaf trichomes, were also reported in this study. Together with the fact that only some of the F_2 families have shown early mortality, it is likely that different individuals of *S. aethnensis* and *S. chrysanthemifolius* have variable levels of genetic incompatibilities and that the extent of hybrid breakdown is dependent on the lineages, as suggested in other studies too (Rosenthal *et al.*, 2005).

While it has been reported that seed set was significantly lower when *S. aethnensis* was the maternal parent (Chapman, Forbes, & Abbott, 2005), this was not apparent in this study. In seeds created from F_1 plants, there was sometimes a big difference between total seed sets (across multiple crossing attempts) from each parents of a cross, and in individual seed sets (single attempt) between the two plants of the same cross. The discrepancies could be due to the capitula being at different

flowering stages at the time of crossing, or could point to another manifestation of hybrid breakdown. However, maternal background of these F_1 hybrids were not tested hence no conclusions can be drawn regarding species' influence on seed set.

5.5.2 Leaf architecture, floral and vegetative characteristics in F_2 hybrids

Although each F_2 family showed different signs of hybrid breakdown (for example, family 193B had significantly higher number of effective branches and B71 had significantly smaller leaf area than the rest), there is not one trait that could clearly distinguish one family from another. Unlike *S. aethnensis* and *S. chrysanthemifolius*, F_2 families have no distinctive leaf shapes – they all have variable, intermediate degrees of leaf dissection even within the same family. The control of leaf shape is well-known to involve complex interactions among developmental genes that influence cell division and expansion in various parts of a leaf, not to mention environmental factors and so on (Gonzalez, Vanhaeren, & Inzé, 2012; Tsukaya, 2005, 2006). As almost all leaf traits are correlated in both F_2 hybrids and *S. squalidus*, except for comparisons with compactness (which is estimated from leaf perimeter and area) in F_2 plants, the network of genes underlying the formation of *Senecio* leaves could be linked to some extent. Pleiotropic genes underlying leaf shape (as shown in studies on *Glycine max* (soybean) leaves; Jeong *et al.*, 2012; Sayama *et al.*, 2017) could also play a role in the correlations of leaf traits; but if they are, there would be more than one (otherwise, leaf traits among F_2 families would not be significant different). Another thing to note is that unlike *S. aethnensis* and *S. chrysanthemifolius*, leaf dissection (measured as leaf perimeter to area ratio in this study) is very variable within each F_2 family and among *S. squalidus*, and cannot be used as an obvious trait to distinguish between groups.

In contrast, only one pair of floral traits (number of capitula and apical capitulum pedicel length) in F_2 hybrids were correlated, suggesting floral traits might be controlled by more unlinked genes or non-pleiotropic interactions among genes. Another trait indirectly related to flowering is the number of effective branches, as more branches would mean more stems to bear capitula, hence

potentially higher reproductive fitness. However, more growth in one trait does not necessarily mean the same in another, as observed in F_2 hybrids: stem diameter was negatively correlated with apical capitulum diameter; whereas plant height was not significantly correlated with the number of effective branches.

5.5.3 Nucleocytoplasmic effect on F_2 hybrids

The two F_2 families with reciprocal F_0 parents showed significant differences overall and in four traits. This is not surprising, as nucleocytoplasmic differences have been proposed in this system intrinsically (Brennan *et al.*, 2014); whereas this study showed that nucleocytoplasmic differences could manifest in phenotypes of *Senecio* as well, as shown in other systems (reviewed in Levin, 2003). It is interesting that there was a lack of significant differences in floral traits in the reciprocal F_2 hybrids; this could suggest that nucleocytoplasmic differences would not result in different floral developments. However, this should be treated with caution as only four floral traits were analysed.

Only two studies (Brennan *et al.*, 2014; current study) have investigated the effect of nucleocytoplasmic incompatibilities in this system so far. Hence more research is needed to study the effects in detail and to distinguish between the extent of hybrid breakdown resulted from nucleocytoplasmic incompatibilities and genetic incompatibilities between the two Etnean *Senecio* species. One interesting future direction is to study the effect in *S. squalidus* in the U.K. through analyses of species-specific (to *S. aethnensis* and *S. chrysanthemifolius*) nuclear and cytoplasmic genes. By comparing the distribution of each species' contribution to the nucleocytoplasmic background of *S. squalidus* through space and time (using herbarium collections available), insights could be gained on the benefits each cytoplasmic background have on different functions related to fitness and reproduction. This can build on previous work that investigated the proportion of *S. chrysanthemifolius* ancestry in *S. squalidus* samples from different locations (Abbott *et al.* 2009).

5.5.4 Comparing natural and synthetic hybrids

It is clear that *S. squalidus* and synthetic F_2 hybrids between *S. aethnensis* and *S. chrysanthemifolius* are different even when both were grown in the same greenhouse setting, as all the traits measure are significantly different between *S. squalidus* and F_2 hybrids, and between greenhouse-grown *S. squalidus and* F_2 hybrids, except that the leaf perimeter to area ratio were not significantly different in the latter. This is not surprising, as *S. squalidus* had surely undergone many more generations since hybridisation, compared to F_2 hybrids, while they accumulate different changes to their phenotypes and genotypes. The differences and changes could come from introgression from other native *Senecio* species in the UK, such as *S. vulgaris* and *S. viscosus,* as reported in literature (Hegarty *et al.*, 2008; Walter *et al.*, 2020a). Interestingly, *S. squalidus* have smaller but wider leaves than F_2 hybrids. The fact that *S. squalidus* and F_2 plants showed significant differences adds to existing literature that revealed hybrid species had undergone many phenotypic and genetic changes after speciation (Rieseberg, 2006; Rosenthal *et al.*, 2005; Zhang *et al.*, 2016). It is fairly clear that *S. squalidus* showed some morphological novelty compared to synthetic hybrids and certain novelty in terms of adaptation to the habitat in the U.K. in poor soils such as train tracks and stone walls. Future research should focus on the genetic aspects of adaptation in order to pinpoint the underlying causes of these novelties.

On the other hand, in the greenhouse setting, both groups of plants have similar perimeter to area ratio, pointing to some phenotypic plasticity within *S. squalidus*, further discussed below.

5.5.5 Effect of synthetic environments on phenotypes

Environmental conditions seem to have an effect on the phenotypes of *Senecio* plants, as wild- and greenhouse-grown *S. squalidus* were shown to be significantly different overall. Typically, one would expect plants to grow better in the greenhouse as the environment is less harsh than the wild – there is abundant light and water, rich soil, almost no predation at all, and so on. Hence it is not surprising that *S. squalidus* plants with the same genetic origins showed significantly different phenotypes in the greenhouse setting. However, they did not show significantly different leaf width,

area and perimeter to area ratio. Perimeter to area ratio seems to be a consistent trait between plants in the two settings and would continue to serve as a good proxy for leaf dissection, as it has been used in *Chapter 3* (Wong *et al.*, 2020). However, one should note that leaf dissection is more than merely this ratio, as there could be other factors that describe it such as the angle, number and size of lobes.

The fact that leaves are significantly longer but not wider or bigger in terms of area is interesting. However, no one simple explanation could be used to explain these results as many developmental biologists still try to understand leaf morphogenesis, which are controlled by multiple genes underlying different cell functions. Many leaf developmental studies (e.g. Gonzalez *et al.*, 2012; Tsukaya, 2005, 2006) utilise *Arabidopsis thaliana*. However, recent studies have incorporated *A. thaliana*'s close relatives which have dissected leaves and/ or leaflets to study morphogenesis, such as *Cardamine hirsuta*. For instance, recent studies have shown that leaf form of *C. hirsutai* progressed in an age-dependent fashioin (heterochrony; Cartolano *et al.*, 2015). The homeobox gene *REDUCED COMPLEXITY* (*RCO*) was also found to sculpt leaf shapes – it is a product of gene duplication in Brassicaceae and subsequently lost in *A. thaliana*, hence it only has simple leaves (Vlad *et al.*, 2014). Apart from *RCO*, Kierzkowski *et al.* (2019) studied another homeobox gene, *SHOOTMERISTEMLESS (STM)*, that is involved in shape diversity and successfully recreated *C. hirsuta*'s complex leaf morphology in *A. thaliana* leaves. It remains unclear whether these promising findings would apply to species from other families and genera, such as *Senecio*. It will require a lot of future work to explore the molecular basis of leaf shape evolution in *Senecio*.

5.5.6 Implications for hybrid reproductivity and viability on Mount Etna

Based on results of this crossing experiment and previous studies (Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009), it is unequivocal that hybrid breakdown exists in the hybrids between *S. aethnensis* and *S. chrysanthemifolius*. However, the extent and types of hybrid breakdown seem to be dependent on the genetic background. This may help explain the prevalence of hybrids plants in the natural hybrid zone on Mount Etna – *S. aethnensis* and *S. chrysanthemifolius*

that have less incompatibilities could create hybrids and subsequent offspring that perform relatively better than the ones with more severe hybrid breakdown. So far, only the study of Hegarty *et al.* (2009) has investigated hybrids past the second generation. Hence, more evidence is needed to see whether hybrid fitness could restore in later generations.

Together with the findings in *Chapter 3* (Wong *et al.*, 2020) and previously reported intrinsic incompatibilities between *S. aethnensis* and *S. chrysanthemifolius* (e.g. Brennan *et al.*, 2014; Chapman *et al.*, 2016; Osborne *et al.*, 2016), hybrid breakdown that would result in low germination, survival and reproductive success adds to the extrinsic and intrinsic selection to keep the number of hybrids down on Mount Etna, and thus help maintaining the divergence of the two parental species.

Generating four new F_2 families, this study has dramatically increased the number of synthetic hybrid lineages studied (only four previously, one in each of Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009; Ross, 2010) in this system. It has been confirmed that hybrid breakdown does occur in hybrids of *S. aethnensis* and *S. chrysanthemifolius* and its extent depends on its lineage. Previously unidentified signs of hybrid breakdown were recorded, including the presence of leaf trichomes and capitula lacking ray flowers. Comparisons between offspring derived from a reciprocal cross also reported for the first time morphological differences that stem from nucleocytoplasmic incompatibilities. Further investigation should be done to locate the genes responsible for these incompatibilities, with reference to the differences in survival and traits identified in this study. Regarding the natural hybrid, *S. squalidus* in the U.K., it showed some phenotypic plasticity under different growing conditions and is significantly different from synthetic hybrids in leaf traits. More research should be done involving genomic sequencing and more morphological traits in order to further study the amount of novelty *S. squalidus* has acquired since its speciation, with the help of other native *Senecio* species to identify potential introgression that assisted in adaptation and establishment. Future research can also focus on the complex interplay among developmental and genetic processes that determine the leaf shape in this system.

6. General Discussion

Speciation in the Etnean *Senecio* system has been explored using a variety of data previously, such as phenotypic traits (Ross, 2010; Walter *et al.*, 2020b), PCR markers (Brennan *et al.*, 2009, 2014, 2016, 2019; Ross, 2010), transcriptomes (Chapman *et al.*, 2013, 2016; Muir *et al.*, 2013; Osborne *et al.*, 2013; Walter *et al.*, 2020b), and reduced-representation genomes (Filatov *et al.*, 2016; Wong *et al.*, 2020). To add to the existing literature, this thesis presented analyses using reduced-representation genomic data for *Senecio* populations at intermediate elevations for demographic modelling and cline analyses, for the first time genome sequences of *S. aethnensis* and *S. chrysanthemifolius* in genome scans for positive selection, and phenotypic comparisons between F_2 hybrids between the two species and *S. squalidus* collected in the UK. In this final chapter, I will summarise the findings presented in this thesis and discuss future directions in this system.

6.1 Demographic history and divergence in Etnean Senecio

Demographic modelling in *Chapter 3* revealed several crucial features during the speciation of *S. aethnensis* and *S. chrysanthemifolius*, including heterogenous gene flow, bidirectional gene flow and population size change in the past. This study also found a secondary contact scenario and a gene flow since divergence scenario had almost equal likelihood [whereas, Filatov *et al.* (2016) concluded a secondary scenario is unlikely, though their models did not incorporate demographic features mentioned above]. Either way is possible, and allopatric or secondary contact phases could have happened in the latter scenario but were too short to be reflected in genomic data. Previous young estimates of the age of hybrid zone to be fewer than 100 years (Brennan *et al.*, 2009) could be seen as support of the secondary contact scenario. However, this recent secondary contact scenario would create strong linkage disequilibrium among loci, which was not observed in *S. aethnensis* and *S. chrysanthemifolius*. It is likely that gene flow had been happening for much longer to generate the observed genetic similarities and differentiation patterns. It is probable that volcanic activities have

regularly eliminated hybrid zones, or the hybrid zones had been maintained by early generation hybrids due to reproductive issues in later generations. Research regarding fitness of later generation hybrids are yet to be done. If it is the former case, an interesting question arises: did volcanic activities that regularly destroy hybrid zones keep the two species from species fusion? This would require large-scale experiments to recreate a hybrid zone that involves many generations to quantify the amount of hybrid breakdown, level of backcrosses, and to simulate the removal of hybrid zone.

Of course, it would be interesting to know the geographic mode that initiated speciation. However, one might never find that out without historical evidence. Focusing on this would also lead to ignorance of important processes that happened during speciation (Fitzpatrick, Fordyce, & Gavrilets, 2009; Gavrilets, 2003). Hence, it is more important to focus on the processes in the speciation continuum, as the rest of the chapters in this thesis did.

One rather unexplored aspect of speciation in the system is chromosomal rearrangements. Chromosomal rearrangements have long been suggested to have a role in speciation and generation of reproductive isolation (e.g. Hoffmann & Rieseberg, 2008; Noor & Feder, 2006; Rieseberg, Whitton, & Gardner, 1999). It has been speculated for the Etnean *Senecio* system presented here, although it was not confirmed due to lack of research and lack of resolution in linkage maps (Brennan *et al.*, 2014). Hence, future research could focus on generating high-density linkage maps to investigate potential structural rearrangements between *S. aethnensis* and *S. chrysanthemifolius*. On the other hand, the pollination syndrome, self incompatibility, has been studied in the hybrid species, *S. squalidus* (e.g. Allen, Thorogood, Hegarty, Lexer, & Hiscock, 2011; Hiscock, 2000), but not in the two Etnean species. However, it has been suggested that self incompatibility also contributes to intra and inter-specific incompatibilities (Rieseberg & Willis, 2007), and had been studied in genera such as *Nicotiana* (Murfett *et al.*, 1996) and *Lycoperscon* (Bernacchi & Tanksley, 1997). Future studies on the identity and genomic position of the self-incompatibility (*S*) locus or loci in the Etnean species, as well as the extent of its role in reproductive isolation would be interesting.

6.2 Selective mechanisms in Etnean *Senecio*

Chapter 3 and *4* have shown that divergent selection and positive selection on adaptation and speciation genes likely contributed in shaping the Etnean *Senecio* system. Positive selection is also very widespread within and across linkage groups, with various strengths and ages. Selection likely acts with the previously identified intrinsic incompatibilities in generating and maintaining divergence between the two species. Although the bottom-up approaches employed in the two chapters have uncovered new candidate genes and information about the extent of selection in the genome, one drawback of these approaches, especially for non-model organisms, is that it is difficult to know whether the markers which did not have gene annotations are underlying adaptive traits or whether they contribute to intrinsic post-zygotic isolation only. If they are not associated with adaptive traits, are these intrinsic selection (and the intrinsic incompatibilities revealed in previous studies; Brennan *et al.*, 2014; Chapman *et al.*, 2016) caused by genetic drift or close proximity to adaptation genes? To answer this, a more robust genome and genomic resources would be needed.

Chapter 3 and *4, t*ogether with the differentially expressed genes identified in Chapman *et al.* (2013), have now provided a long list of candidate genes that might have contributed to speciation and adaptation, including but not limited to those related to metal ions, UV, water deprivation, photosynthesis, floral development, defence response and growth. As all the candidate genes are inferred from genetic data and based on annotations made for other model species, functional validation on these genes is necessary to confirm if they actually contributed to adaptative and reproductive isolation (Wolf & Ellegren, 2017). This would require tremendous efforts to disentangle the relationship between adaptation traits and their underlying genes, as adaptation traits likely have complex genetic background (Hendry, 2009; Weedon & Frayling, 2008; Wolf *et al.*, 2010). Also, a crucial next step is to quantify the change in environmental conditions along the elevational gradient. This can be used to confirm whether the extrinsic factors related to the above genes actually showed differences at high and low elevations. The clines visualising the transition in the environmental conditions could also be compared with the clines of the genetic markers and see if they match in shape and position, as a match could mean correlation between the two. The environmental aspect of this speciation system is largely unexplored partly due to logistic reasons (such as the field site being

a popular tourist destination), although it would provide useful insights into the missing part in environment-phenotype-genotype interactions.

Physiological (with the exception of ecophysiology) and biochemical basis of speciation and adaptation are unexplored in this system. However, they could also have a role in speciation and adaptation. Genes that involve cellular processes and organelles in *Chapter 4* might be candidates for these bases and understanding these processes would be a great addition to the body of knowledge in this system alongside what we know about the species' morphology, genomics and environmental conditions. Questions could be asked regarding the position of genes and their functions in the chains of protein networks. For instance, it has been suggested that hybrid breakdown phenotypes such as necrosis are similar to those of pathogen attacks, hinting that hybrid breakdown might be a consequence of changes in genes related to pathogen resistance (Rieseberg & Willis, 2007). Indeed, numerous genes related various defence responses have been identified in both highly differentiated genes in *Chapter 3* and putative selective sweeps in *Chapter 4*. Experiments could be done to knock out well-known pathogen response genes to test this hypothesis.

6.3 Hybrids, hybridisation and maintenance of the hybrid zone

Chapter 5 of this thesis showed hybrid breakdown phenotypes in all four F_2 families studied, while previous studies have documented both the presence (Brennan *et al.*, 2014; Chapman *et al.*, 2016; Hegarty *et al.*, 2009) and absence (Ross, 2010) of hybrid breakdown in F_2 hybrids between *S. aethnensis* and *S. chrysanthemifolius*. Some of the previously reported hybrid breakdown phenotypes, such as necrosis and albinism, were not found in the work reported in *Chapter 5*; although new breakdown traits, such as presence of leaf trichomes, have been found. As this is the first study to record hybrid breakdown in multiple F_2 families at the same time under the same protocol, it can be concluded that the types and extent of hybrid breakdown likely depends on genetic background of the parents and their level of intrinsic incompatibility, without artefacts created by different experimental settings in previous studies. This also supports the idea that reproductive isolation between the two Etnean species is not complete, although so much hybrid breakdown has already

evolved in fewer than 200,000 years. Given each F_2 family exhibit a wide range of leaf shapes from non-dissected to highly-dissected whereas hybrids on Mount Etna show a transition in leaf dissection along the elevational gradient, what is the adaptive effect of leaf shape and what is the selective agent on Mount Etna? Clearly, hybrids with the 'maladapted' leaf dissections at different elevations had been selected against and that would suggest very high selection pressure based on leaf shape alone. This remains a mystery, as Ross (2010) found inconsistent associations between leaf shape and performance in the field, as well as a lack of effect of leaf dissection on other ecophysiological traits. The only thing certain is that leaf dissection appear to have an adaptive function.

Chapter 5 also revealed significant phenotypic differences between synthetic hybrids grown in the greenhouse and the natural hybrid species *S. squalidus* collected in the field. This is not surprising, as *S. squalidus* has presumably undergone more generations and hence accumulated more genetic novelty through mutations and/ or introgression from other native UK species. This also supports the findings of Ross (2010), who showed *S. squalidus* outperformed Etnean hybrids in common garden experiments, suggesting post-invasion evolutionary changes in the former. While *Chapter 5* and Ross (2010) both found significant phenotypic differences between *S. squalidus* and synthetic hybrids, *Chapter 5* added to the picture by accounting for the potential phenotypic plasticity of *S. squalidus* and confirmed significant differences in a few leaf traits between field-collected and greenhouse grown *S. squalidus* of the same genetic background. This would imply certain amount of phenotypic 'novelty' found in *S. squalidus* is due to environmental differences instead of permanent genetic changes. This is a caveat future research that focuses on the evolutionary changes of *S. squalidus* after colonising the UK has to bear in mind. This also points to the importance of incorporating environmental data into speciaiton studies as mentioned above. It would provide more information on how the hybrid zone on Mount Etna is maintained, alongisde selection against hybrids idenntified in *Chapter 3*.

There is no doubt hybrids and hybrid zones offer invaluable tools to study speciation. Another promising future research topic is the study of hybrid mis-expression. Gene expression divergence and functional polymorphisms in genes have been shown to be relevant to speciation in previous studies in *Senecio* and other systems (Osborne *et al.*, 2016; reviewed in Wolf *et al.*, 2010). Although quantitative trait loci mapping using transcriptomic data had been carried out in synthetic hybrids in

this system before (Chapman *et al.*, 2013, 2016), no one has examined gene expression and mis-expression patterns in *Senecio* hybrids from the mountain. With the availability of sequencing and analytical tools nowadays, the genomes and transcriptomes of wild hybrids can be sequenced and directly compared to study the interplay between polymorphism and expression divergence at the same time. This would also allow analyses of selection on gene expression in hybrids, and to further examine intrinsic incompatibilities not just in protein-coding regions in transcriptomes (Chapman *et al.*, 2016) and genomic regions sequenced with PCR markers (Brennan *et al.*, 2014), but also at the genomic level in non-coding regions such as regulatory variations.

Compared to its Etnean counterparts, the UK hybrid species, *S. squalidus*, has received less research attention. So far, research has been done focusing on the self-incompatibility syndrome (e.g. Allen *et al.*, 2011; Hiscock, 2000), genetic differentiation from the two Etnean species (e.g. Abbott, James, Forbes, & Comes, 2002; James & Abbott, 2005), hybridisation with native UK species (e.g. Hegarty, Abbott, & Hiscock, 2012; Walter *et al.*, 2020a), and expression divergence among various UK localities (unpublished). Future research should consider the use of genomic sequencing. Potential research topics could include polymorphism patterns among localities, structural changes during homoploid hybrid speciation, novel genotypes and phenotypes that are related to adaptation in the UK, and so on. High-density quantitative trait loci mapping could also be done using hybrid crosses and their frozen tissues generated in the studies of this thesis. Similar to previous linkage mapping studies, previous quantitative trait loci maps suffer from lack of resolution due to the number of markers used. However, frozen tissues can be used for DNA and RNA sequencing and used with traits in *Chapter 5* to generate a denser map. Besides, reproductive success of F_2 families could be estimated based on seed counts and germination rate in subsequent generations. F_2 families that were generated but not germinated here could be added to the pool of hybrids in the analyses to strength the record of potential hybrid breakdown and further establish the link between genetic background and hybrid phenotypes. Later generations using seeds from the unused families can also be generated to see if they would recover from incompatibilities and hybrid breakdown (Hegarty *et al.*, 2009).

6.4 Conclusion - Etnean *Senecio* as a model to study speciation with gene flow

Studies in this thesis were set out to gain more in-depth understanding of adaptation and speciation between *S. aethnensis* and *S. chrysanthemifolius*, and their homoploid hybrid species. Genomic data of hybrids at intermediate elevations on Mount Etna and genomic sequences of the two Etnean species have been analysed for the first time in this system. Through analyses on demography, clines, selection and phenotypes, more questions have been answered but a lot more are yet to be explored. This thesis has shed light on the role of extrinsic and positive selection on speciation and the values of hybrids zones in speciation studies. Now, it is certain that speciation and adaptation in this system involves multiple selective agents; and they are associated with many intrinsic and extrinsic mechanisms that may or may not be manifested in the phenotypes. Hopefully, results in this thesis would help evolutionary biologist understand more about speciation with gene flow and inform processes in other similar systems. As both the Etnean speciation (< 200 mya) and its homoploid hybrid speciation (~ 300 ya) represent two of the very few young systems known to evolutionary biologists that are still in the process of evolving and building up reproductive isolation, they are especially informative for evolutionary biologists to study processes in early stages of speciation and adaptation. It is believed that research in these emerging systems will continue to thrive and contribute to the field of speciation, adaptation and evolutionary biology.

7. References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., … & Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, **26**(2), 229–46. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Abbott, R. J., & Brennan, A. C. (2014). Altitudinal gradients, plant hybrid zones and evolutionary novelty. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **369:** 20130346. https://doi.org/10.1098/rstb.2013.0346

Abbott, R. J., Brennan, A. C., James, J. K., & Forbes, D. G. (2009). Recent hybrid origin and invasion of the British Isles by a self-incompatible species, Oxford ragwort (*Senecio squalidus* L., Asteraceae). *Biological Invasions*, **11**, 1145. https://doi.org/10.1007/s10530-008-9382-3

Abbott, R. J., Hegarty, M. J., Hiscock, S. J., & Brennan, A. C. (2010). Homoploid hybrid speciation in action. *Taxon*, **59**(5), 1375–1396. https://doi.org/10.1002/tax.595005

Abbott, R. J., James, J. K., Forbes, D. G., & Comes, H. P. (2002). Hybrid origin of the Oxford Ragwort, *Senecio squalidus* L.: morphological and allozyme differences between *S. squalidus* and *S. rupestris* Waldst. and Kit. *Watsonia*, **23**, 123–138.

Alexandrino, J., Baird, S. J., Lawson, L., Macey, R. J., Moritz, C., & Wake, D. B. (2005). Strong selection against hybrids at a hybrid zone in the Ensatina ring species complex and its evolutionary implications. *Evolution*, **59**(6), 1334– 1347. https://doi.org/10.1554/04-156

Akashi, H., Osada, N., & Ohta, T. (2012). Weak Selection and Protein Evolution. *Genetics*, **192**, 15–31. http://doi.org/10.1534/genetics.112.140178

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, **1**, 79. http://doi.org/10.1038/s42003-018-0085-8

Alachiotis, N., Stamatakis, A., & Pavlidis, P. (2012) OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, **28**(17), 2274–5. https://doi.org/10.1093/bioinformatics/bts419

Allen, A., Thorogood, C., Hegarty, M., Lexer, C., & Hiscock, S. (2011). Pollen-pistil interactions and self-incompatibility in the Asteraceae: new insights from studies of *Senecio squalidus* (Oxford ragwort). *Annals of Botany*, **108**(4), 687–98. https://doi.org/10.1093/aob/mcr147

Anderson, & Stebbins. (1954). Hybridization as an Evolutionary Stimulus. *Evolution*, **8**(4), 378. http://doi.org/10.2307/2405784

Antonovics, J., & Bradshaw, A. D. (1970). Evolution in closely adjacent plant populations VIII. Clinal patterns at a mine boundary. *Heredity*, **25**(3), 349– 362. https://doi.org/10.1038/hdy.1970.36

Arnold, M. L, & Bennett, B. D. (1993). Natural selection in Louisiana irises: genetic variation and ecological determinants. In: Hybrid Zones and the Evolutionary Process, pp. 115–139. Oxford University Press, Oxford.

Aquadro, C. F., Begun, D. J., & Kindahl, E. C. (1994). Selection, recombination, and DNA polymorphism in Drosophila. In: Non-neutral evolution. pp. 46-56. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-2383-3_4

Bachtrog, D. (2004). Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. *Nature Genetics*, **36**, 518–522. http://doi.org/10.1038/ng1347

Backhaus, A., Kuwabara, A., Bauch, M., Monk, N., Sanguinetti, G., & Fleming, A. (2010). leafprocessor: a new leaf phenotyping tool using contour bending energy and shape cluster analysis. *New Phytologist*, **187**, 251–261. http://doi.org/10.1111/j.1469-8137.2010.03266.x

Bailey, R. I., Tesaker, M. R., Trier, C. N., & Sætre, G.-P. (2015). Strong selection on male plumage in a hybrid zone between a hybrid bird species and one of its parents. *Journal of Evolutionary Biology*, **28**(6), 1257– 1269. https://doi.org/10.1111/jeb.12652

Baldassarre, D. T., White, T. A., Karubian, J., & Webster, M. S. (2014). Genomic and morphological analysis of a semipermeable avian hybrid zone suggests asymmetrical introgression of a sexual signal. *Evolution*, **68**(9), 2644– 2657. https://doi.org/10.1111/evo.12457

Barton, N. H., & Gale, K. S. (1993). *Genetic analysis of hybrid zones*. (pp. 13– 45). New York, NY: Oxford University Press.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**(1), 113– 148. https://doi.org/10.1146/annurev.es.16.110185.000553

Begun, D., & Aquadro, C. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520. https://doi.org/10.1038/356519a0

Bernacchi, D., & Tanksley, S. D. (1997). An interspecific backcross of *Lycopersicon esculentum* x *L. hirsutum*: linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics*, **147**(2), 861–77.

Bohonak, A. J. (2002). IBD (isolation by distance): A program for analyses of isolation by distance. *Journal of Heredity*, **93**, 153– 154. https://doi.org/10.1093/jhered/93.2.153

Bohra, A., Jha, U. C., Adhimoolam, P., Bisht, D., & Singh, N. P. (2016). Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. *Plant Cell Reports*, **35**(5), 967-993. http://doi.org/10.1007/s00299-016-1949-3

Bolger, A., Lohse, M., & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolnick, D., & Fitzpatrick, B. (2007). Sympatric Speciation: Models and Empirical Evidence. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 459–487. http://doi.org/10.1146/annurev.ecolsys.38.091206.095804

Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in R with the ade4 package. *Journal of Statistical Software*, **86**(1), 1– 17. https://doi.org/10.18637/jss.v086.i01

Braverman, J., Hudson, R., Kaplan, N., Langley, C., & Stephan, W. (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**(2), 783–96.

Brelsford, A., & Irwin, D. (2009). Incipient speciation despite little assortative mating: The Yellow-rumped Warbler hybrid zone. *Evolution*, **63**(12), 3050– 3060. https://doi.org/10.1111/j.1558-5646.2009.00777.x

Brennan, A. C., Bridle, J. R., Wang, A.-L., Hiscock, S. J., & Abbott, R. J. (2009). Adaptation and selection in the *Senecio* (Asteraceae) hybrid zone on Mount Etna, Sicily. *New Phytologist*, **183**, 702– 717. https://doi.org/10.1111/j.1469-8137.2009.02944.x

Brennan, A. C., Hiscock, S. J., & Abbott, R. J. (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**, 195– 204. https://doi.org/10.1038/hdy.2014.14

Brennan, A. C., Hiscock, S. J., & Abbott, R. J. (2016). Genomic architecture of phenotypic divergence between two hybridizing plant species along an elevational gradient. *AoB PLANTS*, **8**, plw022. https://doi.org/10.1093/aobpla/plw022

Brennan, A. C., Hiscock, S. J., & Abbott, R. J. (2019). Completing the hybridization triangle: The inheritance of genetic incompatibilities during homoploid hybrid speciation in ragworts (*Senecio*). *AoB PLANTS*, **11**(1), ply078. https://doi.org/10.1093/aobpla/ply078

Bomblies, K., & Weigel, D. (2007). Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nature Review Genetics*, **8**(5), 382-393. http://doi.org/10.1038/nrg2082

Broad Institute (2019). Picard Toolkit.

Buckley, J., Pashalidou, F., Fischer, M., Widmer, A., Mescher M. C., & De Moraes, C. M. (2019). Divergence in Glucosinolate Profiles between High- and Low-Elevation Populations of *Arabidopsis halleri* Correspond to Variation in Field Herbivory and Herbivore Behavioral Preferences. *International Journal of Molecular Sciences*, **20**(1), 174. https://doi.org/10.3390/ijms20010174

Butlin, R. (2005). Recombination and speciation. *Molecular Ecology*, **14**(9), 2621–2635. http://doi.org/10.1111/j.1365-294X.2005.02617.x

Carneiro, M., Baird, S. J. E., Afonso, S., Ramirez, E., Tarroso, P., Teotónio, H., … & Ferrand, N. (2013). Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. *Molecular Ecology*, **22**(9), 2511– 2525. https://doi.org/10.1111/mec.12272

Carneiro, M., Ferrand, N., & Nachman, M. W. (2009). Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics*, **181**(2), 593-606. https://doi.org/10.1534/genetics.108.096826

Cartolano, M., Pieper, B., Lempe, J., Tattersall, A., Huijser, P., Tresch, A., … Tsiantis M. (2015). Heterochrony underpins natural variation in *Cardamine hirsuta* leaf form. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **112**(33), 10539-44. https://doi.org/10.1073/pnas.1419791112

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). *STACKS*: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171– 182. https://doi.org/10.1534/g3.111.000240

Chapman, M. A., Forbes, D. G., & Abbott, R. J. (2005). Pollen competition among two species of *Senecio* (Asteraceae) that form a hybrid zone on Mount Etna, Sicily. *American Journal of Botany*, **92**, 730– 735. https://doi.org/10.3732/ajb.92.4.730

Chapman, M. A., Hiscock, S. J., & Filatov, D. A. (2013). Genomic divergence during speciation driven by adaptation to altitude. *Molecular Biology and Evolution*, **30**, 2553– 2567. https://doi.org/10.1093/molbev/mst168

Chapman, M. A., Hiscock, S. J., & Filatov, D. A. (2016). The genomic bases of morphological divergence and reproductive isolation driven by ecological speciation in *Senecio* (Asteraceae). *Journal of Evolutionary Biology*, **29**, 98– 113. https://doi.org/10.1111/jeb.12765

Charlesworth B. (2012). The Role of Background Selection in Shaping Patterns of Molecular Evolution and Variation: Evidence from Variability on the *Drosophila* X Chromosome. *Genetics*, **191**(1), 233–246. https://doi.org/10.1534/genetics.111.138073

Charlesworth, D, Charlesworth, B., & Morgan, M. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**(4), 1619–32.

Charlesworth, B., Morgan, M., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**(4), 1289–303.

Chen, J., Huang, Y., Brachi, B., Yun, Q.-Z., Zhang, W., Lu, W., & Sun, H. (2019). Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. *Nature Communications*, **10**, 5230. https://doi.org/10.1038/s41467-019-13128-y

Chessel, D., Dufour, A., & Thioulouse, J. (2004). The ade4 package – I: One-Table methods. *R News*, **4**(1), 5– 10.

Christe, C., Stölting, K., Bresadola, L., Fussi, B., Heinze, B., Wegmann, D., & Lexer, C. (2016). Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F_1 fertility and recurrent gene flow. *Molecular Ecology*, **25**, 2482– 2498. https://doi.org/10.1111/mec.13587

Christe, C., Stölting, K., Paris, M., Fraisse, C., Bierne, N., & Lexer, C. (2017). Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Molecular Ecology*, **26**, 59– 76. https://doi.org/10.1111/mec.13765

Cimino, G., & Toscano, G. (1998). Dissolution of trace metals from lava ash: influence on the composition of rainwater in the Mount Etna volcanic area. *Environmental pollution*, **99**, 389–93. https://doi.org/10.1016/S0269-7491(98)00004-9

Cristaldi, D., Fortuna, C., & Gulino, A. (2013). A photoelectron spectroscopy study of lava stones. *Analytical Methods*, **14**, 3458–3462. https://doi.org/10.1039/C3AY40136G

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daroski, H., & Feder, J. (2007). Host plant and latitude-related diapause variation in *Rhagoletis pomonella*: A test for multifaceted life history adaptation on different stages of diapause development. *Journal of Evolutionary Biology*, **20**, 2101– 2112. https://doi.org/10.1111/j.1420-9101.2007.01435.x

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

DeGiorgio, M., Huber, C., Hubisz, M., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, **32**(12), 1895–7. https://doi.org/10.1093/bioinformatics/btw051

Delmore, K., & Irwin, D. (2014). Hybrid songbirds employ intermediate routes in a migratory divide. *Ecology Letters*, **17**(10), 1211– 1218. https://doi.org/10.1111/ele.12326

Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). hzar: Hybrid zone analysis using an R software package. *Molecular Ecology Resources*, **14**, 652– 663. https://doi.org/10.1111/1755-0998.12209

Doyle, J. J., & Doyle, J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemisty Bulletin*, **19**, 11– 15.

Dray, S., & Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**(4), 1– 20.

Dray, S., Dufour, A., & Chessel, D. (2007). The ade4 package – II: Two-Table and K-Table methods. *R News*, **7**(2), 47– 52. https://doi.org/10.18637/jss.v022

Druce, G. C. (1927). *The Flora of Oxfordshire. A topographical and historical account of the flowering plants and ferns found in the country; with biographical notices of the botanists who have contributed to Oxfordshire botany during the last four centuries*. Oxford: Clarendon Press.

Durrett, R., Buttel, L., & Harrison, R. (2000). Spatial models for hybrid zones. *Heredity*, **84**(1), 9– 19. https://doi.org/10.1046/j.1365-2540.2000.00566.x

Durrett, R., & Schweinsberg, J. (2004). Approximating selective sweeps. *Theoretical Population Biology*, **66**(2), 129-138. https://doi.org/10.1016/j.tpb.2004.04.002

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359– 361. https://doi.org/10.1007/s12686-011-9548-7

Egea-Serrano, A., Hangartner, S., Laurila, A., & Räsänen, K. (2014). Multifarious selection through environmental change: Acidity and predator-mediated adaptive divergence in the moor frog (*Rana arvalis*). *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20133266. https://doi.org/10.1098/rspb.2013.3266

Ellis, A. R., Burchett, W. W., Harrar, S. W., & Bathke, A. C. (2017). Nonparametric inference for multivariate data: the R package npmv. *Journal of Statistical Software*, **76**(4), 1-18. http://doi.org/10.18637/jss.v076.i04

Ellison, C. K., & Burton, R. S. (2006). Disruption of mitochondrial function in interpopulation hybrids of *Tigriopis californicus*. *Evolution*, **60**, 1382– 1391. https://doi.org/10.1554/06-210.1

Ellison, C. K., & Burton, R. S. (2008). Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution*, **62**, 631– 638. https://doi.org/10.1111/j.1558-5646.2007.00305.x

Estep, M., McKain, M., Diaz, D., Zhong, J., Hodge, J. G., Hodkinson, T. R., … & Kellogg, E. A.(2014). Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **111**(42), 15149–15154. https://doi.org/10.1073/pnas.1404177111

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611– 2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Laval, G., & Schneider, S. (2005). *Arlequin* (version 3.0): An integrated software package for site genetics data analysis. *Evolutionary Bioinformatics*, **1**, 117693430500100003. https://doi.org/10.1177/117693430500100003

Fay, J., & Wu, C.-I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**(3), 1405–13.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342– 350. https://doi.org/10.1016/j.tig.2012.03.009

Ferris, K., Barnett, L., Blackman, B., & Willis, J. (2017). The genetic architecture of local adaptation and reproductive isolation in sympatry within the *Mimulus guttatus* species complex. *Molecular Ecology*, **26**(1), 208–224. https://doi.org/10.1111/mec.13763

Fernandes, J. B., Wlodzimierz, P., & Henderson, I. R. (2019). Meiotic recombination within plant centromeres. *Current Opinion in Plant Biology*, **48**, 26-35. https://doi.org/10.1016/j.pbi.2019.02.008

Filatov, D. (2009). Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics*, **25**(23), 3189-3190. https://doi.org/10.1093/bioinformatics/btp572

Filatov, D. A., Osborne, O. G., & Papadopulos, A. S. T. (2016). Demographic history of speciation in a *Senecio* altitudinal hybrid zone on Mount Etna. *Molecular Ecology*, **25**, 2467– 2481. https://doi.org/10.1111/mec.13618

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford: Calrendon Press.

Fishman, L., Kelly, A. J., Morgan, E., & Willis, J. H. (2001). A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. *Genetics*, **159**(4), 1701–16.

Fishman, L., Stathos, A., Beardsley, P., Williams, C., & Hill, J. (2013). Chromosomal rearrangements and the genetics of reproductive barriers in *Mimulus* (monkey flowers). *Evolution*, **67**(9), 2547–2560. https://doi.org/10.1111/evo.12154

Fishman, L., & Willis, J. (2001). Evidence for Dobzhansky-Muller incompatibilities contributing to the sterility of hybrids between *Mimulus guttatus* and *M. nasutus. Evolution*, **55**(10), 1932–1942. http://doi.org/10.1111/j.0014-3820.2001.tb01311.x

Fitzpatrick, B. M., Fordyce, J. A., & Gavrilets, S. (2009). Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology*, **22**(11), 2342–2347. https://doi.org/10.1111/j.1420-9101.2009.01833.x

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, **180**, 977– 993. https://doi.org/10.1534/genetics.108.092221

Fuller, R. C. (2008). Genetic incompatibilities in killfish and the role of environment. *Evolution*, **62**, 3056– 3068. https://doi.org/10.1111/j.1558-5646.2008.00518.x

Gailing, O., & Curtu, A. (2014). Interspecific gene flow and maintenance of species integrity in oaks. *Annals of Forest Research*, **57**(1), 1– 14. https://doi.org/10.15287/afr.2014.171

Gavrilets, S. (2003). Perspective: models of speciation: what have we learned in 40 years? *Evolution*, **57**(10), 2197–2215. https://doi.org/10.1111/j.0014-3820.2003.tb00233.x

Gavrilets, S., & Losos, J. (2009). Adaptive Radiation: Contrasting Theory with Data. *Science*, **323**(5915), 732–737. http://doi.org/10.1126/science.1157966

Giammanco, S., Valenza, M., Pignato, S., & Giammanco, G. (1996). Mg, Mn, Fe, and V concentrations in the ground waters of Mount Etna (Sicily). *Water Research*, **30**(2), 378–386. https://doi.org/10.1016/0043-1354(95)00183-2

Giska, I., Farelo, L., Pimenta, J., Seixas, F. A., Ferreira, M. S., Marques, J. P., & Melo-Ferreira, J. (2019). Introgression drives repeated evolution of winter coat color polymorphism in hares. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **116**(48), 24150–24156. https://doi.org/10.1073/pnas.1910471116

Gompert, Z., Lucas, L., Nice, C., Fordyce, J. A., Buerkle, C. A., & Forister, M. L. (2013). Geographically multifarious phenotypic divergence during speciation. *Ecology and Evolution*, **3**, 595– 613. https://doi.org/10.1002/ece3.445

Gonzalez, N., Vanhaeren, H., & Inzé, D. (2012). Leaf size control: complex coordination of cell division and expansion. *Trends in Plant Science*, **17**(6), 332–340. https://doi.org/10.1016/j.tplants.2012.02.003

Groszmann, M., Greaves, I. K., Albertyn, Z. I., Scofield, G. N., Peacock, W. J., & Dennis, E. S. (2011). Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid

vigor. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences,* **108**(6), 2617-2622. http://doi.org/10.1073/pnas.1019217108

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple sites from multidimensional SNP frequency data. *PLOS Genetics*, **5**, e1000695. https://doi.org/10.1371/journal.pgen.1000695

Harris, S. A. (2002). Introduction of Oxford ragwort, *Senecio squalidus* L. (Asteraceae), to the United Kingdom. *Watsonia*, **24**, 31–43.

Hegarty, M., Abbott, R., & Hiscock, S. (2012). Allopolyploid Speciation in Action: The Origins and Evolution of *Senecio cambrensis*. In: Soltis, P. S. & Soltis, D. E., *Polyploidy and Genome Evolution* (pp. 245–270). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31442-1_13

Hegarty, M. J., Barker, G. L., Brennan, A. C., Edwards, K. J., Abbott, R. J., & Hiscock, S. J. (2008). Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in *Senecio*. *Philosophical Transactions of the Royal Society B*, **363**, 3055-3069. http://doi.org/10.1098/rstb.2008.0080

Hegarty, M. J., Barker, G. L., Brennan, A. C., Edwards, K. J., Abbott, R. J., & Hiscock, S. J. (2009). Extreme changes to gene expression associated with homoploid hybrid speciation. *Molecular Ecology*, **18**, 877–889. http://doi.org/10.1111/j.1365-294X.2008.04054.x

Hendry, A. P. (2009). Ecological speciation! Or the lack thereof? *Canadian Journal of Fisheries and Aquatic Sciences*, **66**(8), 1383–1398. http://doi.org/10.1139/F09-074

Hines, H., Counterman, B., Papa, R., de Moura, P. A., Cardoso, M. Z., Lincares, M., … & McMillan, W. O. (2011). Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **10**B(49), 19666–19671. https://doi.org/10.1073/pnas.1110096108

Hiscock, S. (2000). Self-incompatibility in *Senecio squalidus* L. (Asteraceae). *Annals of Botany*, **108**(4), 181–190. https://doi.org/10.1093/aob/mcr147

Hollander, J., Galindo, J., & Butlin, R. K. (2015). Selection on outlier loci and their association with adaptive phenotypes in *Littorina saxatilis* contact zones. *Journal of Evolutionary Biology*, **28**(2), 328– 337. https://doi.org/10.1111/jeb.12564

Hoffmann, A., & Rieseberg, L. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**, 21–42. http://doi.org/10.1146/annurev.ecolsys.39.110707.173532

Hotz, H., Semlitsch, R., Gutmann, E., Guex, G.-D., & Beerli, P. (1999). Spontaneous heterosis in larval life-history traits of hemiclonal frog hybrids. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **96**(5), 2171–2176. https://doi.org/10.1073/pnas.96.5.2171

Hudson, R., & Kaplan, N. (1995). Deleterious background selection with recombination. *Genetics*, **141**(4), 1605–17.

James, J. K., & Abbott, R. J. (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**(12), 2533– 2547. https://doi.org/10.1554/05-306.1

Jeong, N., Suh, S., Kim, M.-H., Lee, S., Moon, J.-K., Kim, H. S., & Jeong, S.-C. (2012). Ln Is a Key Regulator of Leaflet Shape and Number of Seeds per Pod in Soybean. *The Plant Cell*, **24**, 4807–4818. http://doi.org/10.1105/tpc.112.104968

Johansen-Morris, A. D., & Latta, R. G. (2006). Fitness consequences of hybridiszation between ecotypes of *Avena barbata*: hybrid breakdown, hybrid vigor, and transgressive segregation. *Evolution*, **60**(8), 1585–1595. https://doi.org/10.1554/05-680.1

Johnson, N. A. (2010). Hybrid incompatiblity genes: remnants of a genomic battlefield? *Trends in Genetics*, **26**(7), 317-325. http://doi.org/10.1016/j.tig.2010.04.005

Joly, S., Heenan, P., & Lockhart, P. (2009). A Pleistocene inter-tribal allopolyploidization event precedes the species radiation of *Pachycladon* (Brassicaceae) in New Zealand. *Molecular Phylogenetics and Evolution*, **51**(2), 365–372. https://doi.org/10.1016/j.ympev.2009.02.015

Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., … & Orlando, L. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **111**(52), 18655–18660. https://doi.org/10.1073/pnas.1412627111

Kaplan, N., Hudson, R., & Langley, C. (1989). The “hitchhiking effect” revisited. *Genetics*, **123**(4), 887–99.

Kassambara, A. (2020). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.6.0. https://CRAN.R-project.org/package=rstatix

Kassambara, A., & Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Kawakami, T., Butlin, R. K., Adams, M., Paull, D. J., & Cooper, S. J. (2009). Genetic analysis of a chromosomal hybrid zone in the Australian morabine grasshoppers (Vandiemenella, *Viatica* species group). *Evolution*, **63**(1), 139– 152. https://doi.org/10.1111/j.1558-5646.2008.00526.x

Kawecki, T., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, **7**(12), 1225–1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x

Kent, D. H. (1966). Senecio squalidus L. in the British Isles. 4. The recent spread in Scotland. *Glasgow Nature*, **18**, 407–408.

Kierzkowski, D., Runions, A., Vuolo, F., Strauss, S., Lymbouridou, R., Routier-Kierzkowska, A.-L., ... & Tsiantis, M. (2019). A Growth-Based Framework for Leaf Shape Development and Diversity. *Cell*, **177**, 1405–1418. http://doi.org/ 10.1016/j.cell.2019.05.011

Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, **167**(3), 1513–1524. https://doi.org/10.1534/genetics.103.025387

Kim, Y., & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**(2), 765–77.

Kimura, M. Natural selection as the process of accumulating genetic information in adaptive evolution. (1961). *Genetics Research*, **2**(1), 127-140. https://doi.org/10.1017/S0016672300000616

Kimura, M. Evolutionary Rate at the Molecular Level. (1968). *Nature*, **217**, 624–626. https://doi.org/10.1038/217624a0

Kirkpatrick, M., & Barton, N. H. (1997). The strength of indirect selection on female mating preferences. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **94**(4), 1282–1286. https://doi.org/10.1073/pnas.94.4.1282

Kirkpatrick, M., & Ravigné, V. (2002). Speciation by natural and sexual selection: models and experiments. *The American Naturalist*, **Supp 3**, S22–35. http://doi.org/10.1086/338370

Körner, C. (2007). The use of "altitude" in ecological research. *Trends in Ecology & Evolution*, **22**, 569– 574. https://doi.org/10.1016/j.tree.2007.09.006

Kruuk, L. E., Baird, S. J., Gale, K. S., & Barton, N. H. (1999). A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**(4), 1959– 1971.

Kubota, S., Iwasaki, T., Hanada, K., Nagano, A. J., Fujiyama, A., Toyoda, A., & Morinaga, S.-I. (2015). A Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in a Wild *Arabidopsis* Species. *PLOS Genetics*, **11**(9), e1005361. https://doi.org/10.1371/journal.pgen.1005361

Lampert, K. P., & Schartl, M. (2008). The origin and evolution of a unisexual hybrid: *Poecilia formosa*. *Philosophical Transactions of the Royal Society of London B*, **363**(1505), 2901–9. http://doi.org/10.1098/rstb.2008.0040

Lee, J.-Y., Mummenhoff, K., & Bowman, J. (2002). Allopolyploidization and evolution of species with reduced floral structures in *Lepidium* L. (Brassicaceae). *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **99**(26), 16835–16840. https://doi.org/10.1073/pnas.242415399

Lehnert, S. J., Kess, T., Bentzen, M. P., Kent, M. P., Lien, S., Gilbey, J., & Bradbury, I. R. (2019). Genomic signatures and correlates of widespread population declines in salmon. *Nature Communications*, **10**, 2996. https://doi.org/10.1038/s41467-019-10972-w

Leinonen, P., Remington, D., & Savolainen, O. (2011). Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. *Evolution*, **65**(1), 90–107. https://doi.org/10.1111/j.1558-5646.2010.01119.x

Levin D. A. (2003). The Cytoplasmic Factor in Plant Speciation. *Systematic Botany*, **28**, 5–11.

Lexer, C., Welch, M. E., Durphy, J. L., & Rieseberg, L. H. (2003). Natural selection for salt tolerance quantitative trait loci (QTLs) in wild sunflower hybrids: Implications for the origin of *Helianthus paradoxus*, a diploid hybrid species. *Molecular Ecology*, **12**(5), 1225–1235. https://doi.org/10.1046/j.1365-294X.2003.01803.x

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Y., Costello, J., Holloway, A., & Hahn, M. (2008). "Reverse ecology" and the power of population genomics. *Evolution*, **62**(12), 2984–2994. http://doi.org/10.1111/j.1558-5646.2008.00486.x

Macholán, M., Munclinger, P., Šugerková, M., Dufková, P., Bímová, B., Božíková, E., … & Piálek, J. (2007). Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution*, **61** 4), 746– 771. https://doi.org/10.1111/j.1558-5646.2007.00065.x

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**, 229– 237. https://doi.org/10.1016/j.tree.2005.02.010

Mallet, J., Beltrán, M., Neukirchen, W., & Linares, M. (2007). Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28. https://doi.org/10.1186/1471-2148-7-28

Mallick, S., Gnerre, S., Muller, P., & Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Research*, **19**, 922-933. http://doi.org/10.1101/gr.086512.108

Martin, N., & Willis, J. (2010). Geographical variation in postzygotic isolation and its genetic basis within and between two *Mimulus* species. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 2469–78. http://doi.org/10.1098/rstb.2010.0030

Matos, I., Machado, M. P., Schartl, M., & Coelho, M. M. (2015). Gene Expression Dosage Regulation in an Allopolyploid Fish. *PLOS ONE*, **10**(3), e0116309. https://doi.org/10.1371/journal.pone.0116309

Matsubara, K., Ando, T., Mizubayashi, T., Ito, S., & Yano, M. (2007) Identification and linkage mapping of complementary recessive genes causing hybrid breakdown in an intraspecific rice cross. *Theoretical and Applied Genetics*, **115**, 179–186. http://doi.org/10.1007/s00122-007-0553-x

Mayr, E. (1942). *Systematics and the Origin of Species*. New York: Columbia Univ. Press.

Mayr, E. (1999). Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press.

McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genetics*, **5**(5), e1000471. https://doi.org/10.1371/journal.pgen.1000471

Montejo-Kovacevich, G., Smith, J., Meier, J., Bacquet, C. N., Whiltshire-Romero, E., Nadeau, N. J., & Jiggins, C. D. (2019). Altitude and life-history shape the evolution of *Heliconius* wings. *Evolution*, 73(12), 2436-2450. https://doi.org/10.1111/evo.13865

Mueller, J. (2004). Linkage disequilibrium for different scales and applications. *Briefings in Bioinformatics*, **5**(4), 355– 364. https://doi.org/10.1093/bib/5.4.355

Muir, G., Osborne, O., Sarasa, J., Hiscock, S., & Filatov, D. (2013). Recent ecological selection on regulartory divergence is shaping clinal variation in *Senecio* on Mount Etna. *Evolution*, **67**, 3032– 3042. https://doi.org/10.1111/evo.12157

Murfett, J., Strabala, T. J., Zurek, D., Mou, B., Beecher, B., & McClure, B. A. (1996). S RNase and Interspecific Pollen Rejection in the Genus *Nicotiana*: Multiple Pollen-Rejection Pathways Contribute to Unilateral Incompatibility between Self-Incompatible and Self-Compatible Species. *The Plant Cell*, **8**, 943-958. https://doi.org/10.1105/tpc.8.6.943

Nachman M. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, **17**(9), 481–485. https://doi.org/10.1016/S0168-9525(01)02409-X

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics & Development*, **12**(6), 657-663. https://doi.org/10.1016/S0959-437X(02)00358-1

Nadeau, N., Whibley, A., Jones, R., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., & Jiggins, C. D. (2012). Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**(1587), 343–53. http://doi.org/10.1098/rstb.2011.0198

Ndjiondjop, M., Alachiotis, N., Pavlidis, P., Goungoulou, A., Kpeki, S. B., Zhao, D., & Semagn, K. (2019). Comparisons of molecular diversity indices, selective sweeps and population structure of African rice with its wild progenitor and Asian rice. *Theoretical and Applied Genetics*, **132**, 1145–1158. https://doi.org/10.1007/s00122-018-3268-2

Nevado, B., Contreras-Ortiz, N., Hughes, C., & Filatov, D. A. (2018). Pleistocene glacial cycles drive isolation, gene flow and speciation in the high-elevation Andes. *New Phytologist*, **219**, 779– 793. https://doi.org/10.1111/nph.15243

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566-1575. http://doi.org/10.1101/gr.4252305

Noor, M., & Feder, J. (2006). Speciation genetics: evolving approaches. *Nature Reviews Genetics*, **7**, 851–861. https://doi.org/10.1038/nrg1968

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009a). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**(3), 375– 402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Nosil, P., Harmon, L., & Seehausen, O. (2009b). Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, **24**, 145– 156. https://doi.org/10.1016/j.tree.2008.10.011

Orengo, D., & Aguadé, M. (2004). Detecting the Footprint of Positive Selection in a European Population of *Drosophila melanogaster* Multilocus Pattern of Variation and Distance to Coding Regions. *Genetics*, **167**(4), 1759–1766. https://doi.org/10.1534/genetics.104.028969

Ortego, J., Gugger, P., & Sork, V. (2017). Impacts of human-induced environmental disturbances on hybridization between two ecologically differentiated Californian oak species. *New Phytologist*, **213**, 942– 955. https://doi.org/10.1111/nph.14182

Osborne, O., Batstone, T. E., Hiscock, S. J., & Filatov, D. A. (2013). Rapid speciation with gene flow following the formation of Mount Etna. *Genome Biology and Evolution*, **5**, 1704– 1715. https://doi.org/10.1093/gbe/evt127

Osborne, O., Chapman, M., Nevado, B., & Filatov, D. (2016). Maintenance of Species Boundaries Despite Ongoing Gene Flow in Ragworts. *Genome Biology and Evolution*, **8**, 1038–1047. http://doi.org/10.1093/gbe/evw053

Pavlidis, P., & Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, **24**, 7. https://doi.org/10.1186/s40709-017-0064-0

Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, **30**(9), 2224–2234. https://doi.org/10.1093/molbev/mst112

Le Pennec, G., Butlin, R. K., Jonsson, P. R., Larsson, A. I., Lindborg, J., Bergström, E., … & Johannesson, K. (2017). Adaptation to dislodgement risk on wave-swept rocky shores in the snail *Littorina saxatilis*. *PloS One*, **12**(10), e0186901. https://doi.org/10.1371/journal.pone.0186901

Phillips, B. L., Baird, S. J., & Moritz, C. (2004). When vicars meets: A narrow contact zone between morphologically cryptic phylogeographic lineages of the Rainforest Skink, *Carlia Rubrigularis*. *Evolution*, **58**, 1536– 1548. https://doi.org/10.1111/j.0014-3820.2004.tb01734.x

Plötner, B., Nurmi, M., Fischer, A., Watanabe, M., Schneeberger, K., Holm, S., & Laitinen, R. A. E. (2017). Chlorosis caused by two recessively interacting genes reveals a role of RNA helicase in hybrid breakdown in *Arabidopsis thaliana*. *The Plant Journal*, **91**, 251–262. http://doi.org/10.1111/tpj.13560

Polyakov, A. V., White, T. A., Jones, R. M., Borodin, P. M., & Searle, J. B. (2011). Natural hybridization between extremely divergent chromosomal races of the common shrew (*Sorex araneus*, Soricidae, Soricomorpha): Hybrid zone in Siberia. *Journal of Evolutionary Biology*, **24**(7), 1393– 1402. https://doi.org/10.1111/j.1420-9101.2011.02266.x

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of site structure using multilocus genotype data. *Genetics*, **155**, 945– 959.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rasmann, S., Pellissier, L., Defossez, E., Jactel, H., & Kunstler, G. (2014). Climate-driven change in plant–insect interactions along elevation gradients. *Functional Ecology*, **24**, 46–54. https://doi.org/10.1111/1365-2435.12135

Revelle, W. (2019) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. R package 1.9.12. https://CRAN.R-project.org/package=psych

van Riemsdijk, I., Butlin, R. K., Wielstra, B., & Arntzen, J. W. (2019). Testing an hypothesis of hybrid zone movement for toads in France. *Molecular Ecology*, **28**(5), 1070– 1083. https://doi.org/10.1111/mec.15005

Rieseberg, L. (1997). Hybrid origins of plant species. *Ecology and Systematics*, **28**, 359–389. https://doi.org/10.1146/annurev.ecolsys.28.1.359

Rieseberg, L. (2006). Hybrid speciation in wild sunflowers 1. *Annals of the Missouri Botanical Garden*, **99**, 34–48. https://doi.org/10.3417/0026-6493(2006)93[34:HSIWS]2.0.CO;2

Rieseberg, L. H., Whitton, J., & Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**(2), 713–27.

Rieseberg, L., & Willis, J. (2007). Plant speciation. *Science*, **317**(5840), 910–914. http://doi.org/10.1126/science.1137729

Riihimäki, M., & Savolainen, O. (2004). Environmental and genetic effects on flowering differences between northern and southern populations of *Arabidopsis lyrata* (Brassicaceae). *American Journal of Botany*, **91**(7), 1036–1045.

Roe, A. D., MacQuarrie, C. J. K., Gros-Louis, M.-C., Simpson, J. D., Lamarche, J., Beardmore, T., … & Isabel, N. (2014). Fitness dynamics within a poplar hybrid zone: I. Prezygotic and postzygotic barriers impacting a native poplar hybrid stand. *Ecology and Evolution*, **4**, 1629– 1647. https://doi.org/10.1002/ece3.1029

Rosenthal, D., Rieseberg, L., & Donovan, L. (2005). Re-creating ancient hybrid species' complex phenotypes from early-generation synthetic hybrids: three examples using wild sunflowers. *The American Naturalist*, **166**(1), 26–41. http://doi.org/10.1086/430527

Ross, R. I. (2010). *Local adaptation and adaptive divergence in a hybrid species complex in Senecio* (unpublished doctoral thesis). Univeristy of Oxford.

Roselius, K., Stephan, W., & Städler, T. (2005). The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, **171**(2), 753-763. https://doi.org/10.1534/genetics.105.043877

Rundle, H., & Nosil, P. (2005). Ecological speciation. *Ecology Letters*, **8**(3), 336–352. https://doi.org/10.1111/j.1461-0248.2004.00715.x

Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: Pairing non-invasive sampling with genomics in conservation. *PeerJ*, **3**, e1106. https://doi.org/10.7717/peerj.1106

Sabeti, P., Schaffner, S., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., & Lander, E. S. (2006) Positive Natural Selection in the Human Lineage. *Science*, **312**(5780), 1614–1620. http://doi.org/10.1126/science.1124309

Sayama, T., Tanabata, T., Saruta, M., Yamada, R., Anai, T., Kaga, A., & Ishimoto, M. (2017). Confirmation of the pleiotropic control of leaflet shape and number of seeds per pod by the *Ln* gene in induced soybean mutants. *Breeding Science*, **67**(4), 363–369. http://doi.org/10.1270/jsbbs.16201

Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**(1), 23–35. https://doi.org/10.1017/S0016672300014634

Schluter, D. (2009). Evidence for Ecological Speciation and Its Alternative. *Science*, **323**(5915), 737–741. http://doi.org/10.1126/science.1160006

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, **9**, 671– 675. https://doi.org/10.1038/nmeth.2089

Scotti-Saintagne, C., Mariette, S., Porth, I., Goicoechea, P. G., Barreneche, T., Bodénès, C., … & Kremer, A. (2004). Genome scanning for interspecific differentiation between two closely related Oak Species [*Quercus robur* L. and *Q. petraea* (Matt.) Liebl.]. *Genetics*, **168**, 1615– 1626. https://doi.org/10.1534/genetics.104.026849

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in ecology & evolution*, **19**(4), 198–207. https://doi.org/10.1016/j.tree.2004.01.003

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., … & Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176– 192. https://doi.org/10.1038/nrg3644

senecioDB (undated). *senecioDB: A genomic resource for Senecio species*. Retrieved from http://www.seneciodb.org/wordpress/?page_xml:id=22

Shen, H, He, H., Li, J., Chen, W., Wang, X., Guo, L., Peng, Z., He, G., Zhong, S., Qi, Y., Terzaghi, W., & Deng, X. W. (2012). Genome-wide analysis of DNA methylation and gene expression changes in two *Arabidopsis* ecotypes and their reciprocal hybrids. *Plant Cell*, **24**(3), 875-892. http://doi.org/10.1105/tpc.111.094870

Sibthorp, J. (1794). *Flora Oxoniensis, exhibens plantas in agro Oxoniensi sponte crescentes, secundum systema sexuale distributas*. Oxford: Fletcher and Hanwell.

Singhal, S., & Moritz, C. (2012). Strong selection against hybrids maintains a narrow contact zone between morphologically cryptic lineages in a rainforest lizard. *Evolution*, **66**(5), 1474– 1489. https://doi.org/10.1111/j.1558-5646.2011.01539.x

Smadja, C. M., & Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**(24), 5123–5140. http://doi.org/10.1111/j.1365-294X.2011.05350.x

Sobel, J., Stankowski, S., & Streisfeld, M. (2019). Variation in ecophysiological traits might contribute to ecogeographic isolation and divergence between parapatric ecotypes of *Mimulus aurantiacus*. *Journal of Evolutionary Biology*, **32**, 604– 618. https://doi.org/10.1111/jeb.13442

Soltis, P., & Soltis, D. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, **60**, 561–588. https://doi.org/10.1146/annurev.arplant.043008.092039

Spencer, H. (1864). *The principles of biology*. London, Edinburgh: Williams and Norgate.

Stankowski, S., Sobel, J. M., & Streisfeld, M. A. (2017). Geographic cline analysis as a tool for studying genome-wide variation: A case study of pollinator-mediated divergence in a monkeyflower. *Molecular Ecology*, **26**, 107– 122. https://doi.org/10.1111/mec.13645

Stebbins, G. (1950). Variation and Evolution in Plants. New York: Columbia University Press. https://doi.org/10.7312/steb94536

Stephan, W., Song, Y. S., & Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, **172**(4), 2647-2663. https://doi.org/10.1534/genetics.105.050179

Stephan, W., Wiehe, T. H. E., & Lenz, M. (1992). The effect of strongly selected substitution on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical Population Biology*, **41**(2), 237-254. https://doi.org/10.1016/0040-5809(92)90045-U

Stöck, M., Lamatsch, D., Steinlein, C., Epplen, J. T., Grosse, W.-R., Hock, R., & Schartl, M. (2002). A bisexually reproducing all-triploid vertebrate. *Nature Genetics*, **30**, 325–328. https://doi.org/10.1038/ng839

Szymura, J. M., & Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution*, **40**(6), 1141– 1159. https://doi.org/10.1111/j.1558-5646.1986.tb05740.x

Szymura, J. M., & Barton, N. H. (1991). The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution*, **45**(2), 237– 261. https://doi.org/10.1111/j.1558-5646.1991.tb04400.x

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., … & Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **115**(43), 11006– 11011. https://doi.org/10.1073/pnas.1801832115

Teeter, K. C., Payseur, B. A., Harris, L. W., Bakewell, M. A., Thibodeau, L. M., O'Brien, J. E., … & Tucker, P. K. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Research*, **18**, 67– 76. https://doi.org/10.1101/gr.6757907

Thornton, K. R., & Jensen, J. D. (2007). Controlling the False-Positive Rate in Multilocus Genome Scans for Selection. *Genetics*, **175**, 737-750. http://doi.org/10.1534/genetics.106.064642

Tine, M., Kuhl, H., Gagnaire, P.-A., Louro, B., Desmarais, E., Martins, R. S. T., … & Reinhardt, R. (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770. https://doi.org/10.1038/ncomms6770

Triplett, J., Clark, L., Fisher, A., & Wen, J. (2014). Independent allopolyploidization events preceded speciation in the temperate and tropical woody bamboos. *New Phytologist*, **204**(1), 66–73. http://doi.org/10.1111/nph.12988

Trunschke, J., & Stöcklin, J. (2017). Plasticity of flower longevity in alpine plants is increased in populations from high elevation compared to low elevation populations. *Alpine Botany*, **127**, 41–51. http://doi.org/10.1007/s00035-016-0176-4

Tsukaya, H. (2005). Leaf shape: genetic controls and environmental factors. *The International Journal of Developmental Biology*, **49**, 547–55. http://doi.org/10.1387/ijdb.041921ht

Tsukaya, H. (2006). Mechanism of leaf-shape determination. *Annual Review of Plant Biology*, **57**, 477–496. https://doi.org/10.1146/annurev.arplant.57.032905.105320

Turner, T., Hahn, M., & Nuzhdin, S. (2005). Genomic Islands of Speciation in *Anopheles gambiae*. *PLoS Biology*, **3**(9), e285. https://doi.org/10.1371/journal.pbio.0030285

Ungerer, M., Baird, S., Pan, J., & Rieseberg, L. (1998). Rapid hybrid speciation in wild sunflowers. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **95**(20), 11757–11762. https://doi.org/10.1073/pnas.95.20.11757

Vaid, N., & Laitinen, R. A. E. (2019). Diverse paths to hybrid incompatibility in *Arabidopsis*. *The Plant Journal*, **97**, 199-213. http://doi.org/10.1111/tpj.14061

Vlad, D., Kierzkowski, D., Rast, M.I., Vuolo, F., Ioio, R. D., Galinha, C., … & Tsiantis, M. (2014). Leaf shape evolution through duplication, regulatory diversification, and loss of a homeobox gene. *Science*, **343**, 780-3. http://doi.org/10.1126/science.1248384

Via, S., & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**(19), 4334–45. http://doi.org/10.1111/j.1365-294X.2008.03921.x

Via, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences. Series B, Biological Sciences*, **106**(Suppl 1), 9939– 9946. https://doi.org/10.1073/pnas.0901397106

Vines, T., Dalziel, A., Albert, A., Veen, T., Schulte, P. M., & Schluter, D. (2016). Cline coupling and uncoupling in a stickleback hybrid zone. *Evolution*, **70**(5), 1023–1038. https://doi.org/10.1111/evo.12917

de Vos, M., Kriksunov, K., & Jander, G. (2008). Indole-3-Acetonitrile Production from Indole Glucosinolates Deters Oviposition by *Pieris rapae*. *Plant Physiology*, **146**(3), 916–926. https://doi.org/10.1104/pp.107.112185

Vu, V. Q. (2011). ggbiplot: A ggplot2 based biplot. R package version 0.55. http://github.com/vqv/ggbiplot

Walker, R. (1833). *The flora of Oxfordshire, and its contiguous counties*. Oxford: H. Slatter. https://doi.org/10.5962/bhl.title.51321

Walter, G., Abbott, R., Brennan, A., Bridle, J. R., Chapman, M., Clark, J., … & Hiscock, S. J. (2020a). *Senecio* as a model system for integrating studies of genotype, phenotype and fitness. *New Phytologist*, **226**, 326–344. http://doi: 10.1111/nph.16434

Walter, G., Clark, J., Cristaudo, A., Nevado, B., Catara, S., Paunov, M., … & Bridle, J. R. (2020b). Adaptation to contrasting habitats underlies distinct plastic responses to environmental variation in two closely related *Senecio* species. *bioRxiv*, 2020.01.24.918201. https://doi.org/10.1101/2020.01.24.918201

Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016). Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, **202**(3), 1185-1200. https://doi.org/10.1534/genetics.115.183152

Weedon, M., & Frayling, T. (2008). Reaching new heights: insights into the genetics of human stature. *Trends in Genetics*, **24**(12), 595–603. http://doi.org/10.1016/j.tig.2008.09.006

Wegary, D., Teklewold, A., Prasanna B., Ertiro, B. T., Alachiotis, N., Negera, D., & Semagn, K. (2019). Molecular diversity and selective sweeps in maize inbred lines adapted to African highlands. *Scientific Reports*, **9**, 13490. https://doi.org/10.1038/s41598-019-49861-z

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**(6), 1358-1370. http://doi.org/10.2307/2408641

Westram, A. M., Galindo, J., Alm Rosenblad, M., Grahame, J. W., Panova, M., & Butlin, R. K. (2014). Do the same genes underlie parallel phenotypic divergence in different *Littorina saxatilis* populations? *Molecular Ecology*, **23**(18), 4603– 4616. https://doi.org/10.1111/mec.12883

Whitham, T. G., Morrow, P. A., & Potts, B. M. (1994). Plant hybrid zones as centers of biodiversity: the herbivore community of two endemic Tasmanian eucalypts. *Oecologia*, **97**, 481–490.

Wiehe, T. (1998). The Effect of Selective Sweeps on the Variance of the Allele Distribution of a Linked Multiallele Locus: Hitchhiking of Microsatellites. *Theoretical Population Biology*, **53**(3), 272–283. https://doi.org/10.1006/tpbi.1997.1346

Wielstra, B., Burke, T., Butlin, R. K., Avcı, A., Üzüm, N., Bozkurt, E., … & Arntzen, J. W. (2017). A genomic footprint of hybrid zone movement in crested newts. *Evolution Letters*, **1**(2), 93– 101. https://doi.org/10.1002/evl3.9

Westram, A. M., Rafajlović, M., Chaube, P., Faria, R., Larsson, T., Panova, M., … & Butlin, R. (2018). Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. *Evolution Letters*, **2**(4), 297–309. https://doi.org/10.1002/evl3.74

Willett, C. S. (2010). Potential fitness trade-offs for thermal tolerance in the intertidal copepod *Tigriopis californicus*. *Evolution*, **64**, 2521 – 2534. https://doi.org/10.1111/j.1558-5646.2010.01008.x

Wolf, J. B., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, **18**, 87–100. https://doi.org/10.1038/nrg.2016.133

Wolf, J., Lindell, J., & Backström, N. (2010). Speciation genetics: current status and evolving approaches. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 1717–33. https://doi.org/10.1098/rstb.2010.0023

Wong, E. L. Y., Nevado, B., Osborne, O. G., Papadopulos, A. S. T., Bridle, J. R., Hiscock, S. J., & Filatov, D. A. (2020). Strong divergent selection at multiple loci in two closely related species of ragworts adapted to high and low elevations on Mount Etna. *Molecular Ecology*, **29**(2), 394–412. https://doi.org/10.1111/mec.15319

Wright, S., Foxe, J., DeRose-Wilson, L., Kawabe, A., Looseley, M., Gaut, B. S., & Charlesworth, D. (2006). Testing for Effects of Recombination Rate on Nucleotide Diversity in Natural Populations of *Arabidopsis lyrata*. *Genetics*, **174**(3), 1421–1430. https://doi.org/10.1534/genetics.106.062588

Wu, C. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851– 865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

Young, N. D. (1996). Concordance and discordance: A tale of two hybrid zones in the pacific coast irises (Iridaceae). *American Journal of Botany*, **83**(12), 1623– 1629. https://doi.org/10.1002/j.1537-2197.1996.tb12820.x

Zhang, D., Pan, Q., Tan, C., Zhu, B., Ge, X., Shao, Y., & Li, Z. (2016). Genome-Wide Gene Expressions Respond Differently to A-subgenome Origins in *Brassica napus* Synthetic Hybrids and Natural Allotetraploid. *Frontiers in Plant Science*, **7**, 1508. http://doi.org/10.3389/fpls.2016.01508

Zuellig, M., & Sweigart, A. (2018). Gene duplicates cause hybrid lethality between sympatric species of *Mimulus*. *PLOS Genetics*, **14**(4), e1007130. http://doi.org/10.1371/journal.pgen.1007130