Identification of barriers to gene flow between *Antirrhinum* species

Daniel Michael Richardson

30/09/2022

Thesis submitted for the Degree of Doctor of Philosophy

University of East Anglia
John Innes Centre

Abstract

For natural diversity to persist, there must be mechanisms in place to protect it from the homogenising effects of gene flow. Studies at natural hybrid zones have shown that, where divergent populations meet and exchange genes, genetic loci involved in adaptive population characteristics can resist gene flow. This results in a homogeneous landscape of genomic divergence, with gene flow resistant regions showing elevated divergence compared to other loci. Identification of these divergent loci may inform about the genetic basis of population differentiation, and is therefore a major aim of speciation genomics. However, genomic divergence is inherently noisy, varying due to cryptic population histories and intrinsic genomic factors. Here I introduce the grouping tree scan as a method for summarising between-population diversity across groups of populations. By comparing between-population divergence across the whole genome for many populations simultaneously, this method reduces the noise associated with within-population effects, and provides increased power for detecting divergence signals that may not be detectable through conventional two-way genome scans. Furthermore, because relationships between populations are determined independently of *a priori* assumptions, the approach is resilient to ascertainment bias. I apply this approach to two sympatric subspecies of *Antirrhinum majus* with contrasting flower colours, demonstrating that colour genes alone may be sufficient to facilitate population divergence through epistatic reproductive barriers. I then expand the approach to look at more distantly related species with distinct growth habits, identifying a subset of genomic regions that may underlie reproductive barriers based on adaptation to different environments. Finally, I outline a bioinformatic approach for detecting sRNA-producing genomic inverted repeats, which may not otherwise be detectable through population comparisons. I propose the grouping tree scan as an extension of the genome scan toolkit, expanding the utility of pooled-sequence data for characterising genetic barriers.

Contents

List of figures

List of tables

Acknowledgements

- My supervisor **Enrico Coen**, who supported and guided my progress, and whose wealth of ideas provided a constant supply of fascinating problems to tackle.
- My second supervisor **Tamas Dalmay**, for useful discussions and incisive feedback pertaining to the sRNA work.
- My third supervisor **Simon Moxon**, for guiding and supporting my development as a bioinformatician, and much intellectual contribution to my sRNA analyses.
- **Desmond Bradley**, for being an inspiring and supportive presence since day one, who has helped me in more ways than I can possibly list.
- **Lucy Copsey**, for keeping an eye on me during my occasional forays into plant work.
- **Tingting Li**, for taking the reins of the growth habit project with aplomb, and always being a joy to work with.
- **Matt Couchman**, for keeping the Coen lab’s data under control.
- **Annabel Whibley**, for patient tuition at the start of the project, and continued provision of good counsel throughout.
- **Yongbiao Xue**, **Sihui Zhu**, **Ting Xu**, and others at the Chinese Academy of Sciences, for the wealth of sequence data that this project couldn’t have done without.
- **David Field**, for the endlessly useful SlidingWindows software tool, and for many interesting discussions.
- **Monique Burrus**, **Christophe Thébaud**, and **Christophe Andalo**, for material collection and knowledge contribution from the hybrid zone and beyond.
- **Leighton Folkes**, for many great programming chats, and contributing to the development of my sRNA analyses.
- **Rachel**, for proofreading, moral support, and sharing my fondness for the John Innes Centre’s impressive array of fishes.

Funding

This work was funded through a Norwich Research Park doctoral training studentship from the Biotechnology and Biological Sciences Research Council, with additional funds from the John Innes Centre.

1: Introduction

Life exhibits a wondrous array of natural diversity, all of which exists due to the process of evolution. Evolution creates and disseminates heritable variation across generations. Fundamentally, this variation is created by the processes of mutation and recombination. Its passage, through individuals over time, is then modulated by natural selection (which favours specific variation) and genetic drift (which acts randomly). Together, these processes create diversity in natural forms. How, though, does this diversity persist and grow over time? In this thesis, I will develop experimental methods to study how populations of closely related organisms can retain their specific adaptive identities, even against a background of pervasive gene flow.

When conceptualising evolution, it is typical to work at the population level. Consider a hypothetical population of reproductively compatible organisms. For simplicity, evolution will be assumed to be neutral (that is, no natural selection will be taking place), and recombination will be ignored. Each generation, mutation will give rise to new diversity by creating new alleles, and genetic drift will 'shuffle' the existing diversity by slightly changing frequencies of existing alleles. Now, assume the population undergoes dispersal, with some individuals migrating away from the established population limits, and establishing a new population. Each population will independently undergo divergence, and the distributions of alleles within each population will increasingly differ over time. However, for as long as the populations remain in contact, they can now undergo gene flow - the exchange of genetic material between populations - through hybrid matings. Sharing genes decreases the proportion of allelic differences between populations, but increases the allelic diversity within populations. Estimations from natural populations suggest that only a few migrant individuals per generation are required for populations to maintain their shared identities (Crow and Kimura, 1970, Slatkin, 1987).

This highly simplified scenario illustrates how diversity can be maintained between populations as they undergo dispersal. However, because the homogenising effect of gene flow is much stronger than the diverging effect of mutation, it offers no means by which the two populations could become differentiated. To enable this, the concept must be expanded to include barriers to gene flow.

Reproductive barriers depend upon genetic and environmental factors

Conceptually, the simplest possible barrier to gene flow would simply stop two populations from meeting. If, for example, populations were to become increasingly geographically separated, it would become more difficult for individuals to make contact. This would attenuate (and eventually eliminate) gene flow, and populations would diverge independently. The emergence of impassable landscape features such as oceans, rivers, and mountains may aid in the isolation process, by preventing populations from migrating back from where they came. Such barriers can all be considered environmental in nature; natural selection is not a factor in introducing barriers, and divergence post isolation proceeds through the random processes of mutation and drift. I will define these environmental barriers as factors that prevent gene flow between populations, but are not genetically encoded.

If barriers between populations were solely environmental, then reestablishment of contact would erode the accumulated between-population divergence. It is therefore generally accepted that the divergence process is completed with the introduction of genetic incompatibilities, which prevent gene flow (Coyne and Orr, 2004). I will define these incompatibilities as genetic barriers - genetically encoded elements that restrict gene flow.

Allelic variation at a single locus is generally considered to be insufficient to maintain a barrier to gene flow where populations remain in contact. Such barriers arise from interactions between two or more genes, and correspond to Dobzhansky-Muller incompatibilities (Sweigart and Willis, 2012). The Dobzhansky-

Muller model is a framework that allows deleterious genetic variation to emerge without being detrimental to the originating populations. It requires that distinct alleles of two or more genes arise within two populations. These alleles must not have a negative effect on fitness within their 'native' genetic background. However, gene flow that rearranges alleles from different populations must be deleterious. The formation of barriers to gene flow between populations is thereby mediated by epistasis (*i.e.* non-additive interactions between alleles at different loci).

Epistasis can result in deleterious consequences, but it can also result in increased fitness through coadaptation. First discussed by Dobzhansky (1950), a group of alleles can be considered coadapted if they act together to promote increased fitness, within the context of their population. An example of a coadapted trait is mimicry in the wing patterns of *Heliconius* butterflies. These butterflies show Müllerian mimicry, a phenomenon whereby species that are distasteful or toxic to predators come to adopt identical colour patterns, in order to warn against predation. The effectiveness of mimics depends on the ability of predators to recognise them. If a mimetic butterfly encounters a population of predators that do not recognise its warning pattern, its fitness will be severely penalised. Combinations of alleles involved in faithfully preserving the 'correct' mimetic pattern are maintained, and mixing of alleles is maladaptive (Mallet and Barton, 1989, Choteau *et al.*, 2016). *Heliconius* wing pattern variation is mostly accounted for by four major effect loci (Reed *et al.*, 2011, Martin *et al.*, 2012, Nadeau *et al.*, 2016, Westerman *et al.*, 2018), with distinct patterns evolving through changes in a set of cis-regulatory "modules" (Van Belleghem *et al.*, 2017, Morris *et al.*, 2019). However, in practice, defining the genetic basis of coadaptation is challenging.

I have briefly outlined how environmental and genetic barriers to gene flow can give rise to reproductive barriers between populations. The extent to which each of these factors plays a part in maintaining the barrier depends on the nature of the barrier. Here, I consider two distinct types of reproductive barriers. I hypothesise that these two barriers are reflected in two distinct instances of speciation within the model angiosperm *Antirrhinum*. To introduce these barriers, I will first give an analogy.

Intrinsic epistatic barriers

To illustrate the first barrier, which will be addressed in Chapter 3, consider the example of road use. Standardisation of the direction in which traffic travels is important for the safety of road users, and essential for facilitating beneficial advancements such as traffic control and the development of more practical vehicles. As of 1986, 76 nations mandated driving on the left-hand side of the road, and 166 on the right (The Rule of the Road: An International Guide to History and Practice by Peter Kincaid, 1986). Neither solution is intrinsically more effective than the other, and one can speculate on the range of cultural, colonial, and legal factors involved in establishing the rule of the road in a given nation. However, implementing a given solution necessitates the standardisation of laws that render the other solution inviable. Once chosen, it is not possible to smoothly transition between systems.

Driving on the left or right exhibits three features:

- It represents two equivalent solutions to a shared problem (efficient traffic flow).
- The distinct solutions operate irrespective of environment (driving on left or right work equally well in cold or warm climates).
- Mixed strategies are disadvantageous (switching to driving to the left in a country which drives on the right is not a good idea).

Suppose a biological trait distinguishing two populations exhibited the same features. Where ‘road handedness’ is maintained by the interaction of laws, traffic adaptations, and driver behaviours, there would be a reproductive barrier that is maintained by the interaction of suites of alleles that together encode a viable adaptive solution. The fitness effect imposed by these barriers is environmentally independent, being a product of inherent factors. Seehausen *et al.* (2014) defined such evolutionary barriers as “intrinsic reproductive barriers” – the activity of the barrier arises through the intrinsic interactions of alleles, independently of environmental (extrinsic) factors. Distinct allele combinations may confer high fitness within each population, but breaking these combinations, as in hybrids between populations, imparts a severe fitness penalty. Because these hypothetical allele combinations interact to generate their phenotypes, they can be more formally described as epistatic. I will therefore refer to these barriers as intrinsic epistatic barriers.

Intrinsic epistatic barriers likely underpin the divergence of plumage between two species of Australian woodswallow, the masked woodswallow and the white-browed woodswallow (Peñalba, Peters, and Joseph, 2022). Both species show minimal differentiation at the genomic level, but strikingly different coloured plumage. These plumage differences do not appear to be associated with mate preference. However, woodswallows showing hybrid plumage types are rare. Modelling suggests that, in isolation, populations may have diverged at genes involved in feather development. Later, populations established secondary contact, and underwent extensive gene flow. Most genes flowed freely, as illustrated by the low interspecific genomic divergence, but those involved in plumage differentiation resisted gene flow. It is therefore possible that the two observed plumage types represent equivalent adaptations, with mixed solutions being disfavoured. Similar barriers may be acting between black-coated carrion crows and grey-coated hooded crows, where two genomic loci explain most variation in plumage, and show evidence of resistance to otherwise pervasive gene flow (Knief *et al.*, 2019).

Differentially adaptive barriers

To introduce the second barrier, discussed in Chapter 4, consider the development of clothing in populations of humans occupying extreme environments.

The native peoples of the arctic circle have prospered within some of the harshest environments on Earth by innovating clothing that is adapted to offer advanced protection from arctic conditions. Examples include the Inuit parka, and its maternally adapted variant the amauti (Lincoln, Cooper, and Loovers, 2020). Without developing similar garments, humans would not have been able to continuously inhabit polar regions. Another example is the colonisation of desert regions of North Africa by peoples such as the Bedouin and Tuareg. Whilst much of North Africa is highly fertile, many cultures traffic and transiently inhabit arid desert regions. This is reflected in many traditional garments worn by these groups, which comprise robes for management of heat and eye protection to shelter from sand (Shkolnik *et al.*, 1980). Cultural groups inhabiting extreme climates have originated distinct sets of environmental adaptations. If either group was to adopt some of the clothing of the other, they would be at a significant disadvantage within their native environments.

Here, the type of clothing exhibits the following features:

- A better solution for a given environment
- Distinct solutions for different environments
- Mixed or hybrid strategies could work in an intermediate environment

Populations with distinct traits exhibiting these features would be differentially adapted to distinct ecological conditions, driving the accumulation of genetic variation improving fitness within the environment. Using the terminology of Seehausen *et al.*, (2014), these barriers are largely extrinsic, because their activity is dependent on environmental factors. To avoid making assumptions about the distinction between intrinsic and extrinsic barriers, I will simply refer to these barriers as differentially adaptative barriers.

Differentially adaptive barriers likely underpin the divergence of *Silene dioica* and *Silene latifolia*, two species of campion (Favre, Widmer, and Karrenberg, 2017). *S. dioica* occupies moister and more elevated sites than *S. latifolia*, which occupies lower, drier areas. *Silene* show widespread distributions throughout Europe, and frequently come into contact. However, hybrids are rare in nature, despite performing well in controlled conditions. By transplanting large populations of each species to the other species' preferred habitat, it was observed that the 'foreign' species always showed reduced fitness compared to the native. Hybrid plants were also transplanted; these generally performed less well than the native parental species, but better than the foreign parent. A later comprehensive analysis of 13 characterised reproductive barriers demonstrated that differential adaptation is likely to be the strongest driver of divergence (Karrenberg *et al.*, 2018). Similar adaptive divergence is seen between inter fertile species of *Mimulus*. *M. lewisii* and *M. cardinalis* show preferences for different altitudes, and reduced fitness when transplanted (Ramsey, Bradshaw, and Schemske, 2003, Angert and Schemske, 2005). The mechanisms by which *S. dioica* and *S. latifolia* have adapted to their environments are subtle, and likely underpinned by many adaptations of individually small effect (Gramlich *et al.*, 2022). Similarly, adaptive survival clothing did not arise fully formed, but is the result of countless individual factors pertaining to technology, material availability, and behaviour.

Detecting genetic barrier loci within genomic islands of divergence

I will now consider the means by which the genes underlying a genetic barrier (barrier genes) might be detected. Because selection on genetic barriers reduces flow, barrier genes will exhibit reduced gene flow (as will regions linked to the genetic barriers). Therefore, studies searching for barrier genes have typically aimed to identify signatures of reduced gene flow amongst highly heterogeneous landscapes of genomic divergence between distinct populations undergoing gene flow (Nosil, Funk, and Ortiz-Barrientos, 2009, Ravinet *et al*. 2017). These signatures are often termed genomic islands of divergence. To directly test for genomic islands of divergence within populations, a means of summarising patterns of genetic variation is required. The field of genomics is generally concerned with characterising genetic variation at the level of genomic sites, or nucleotides. At each site, genetic diversity can be summarised by recording the frequencies of different alleles. For example, consider a single genomic site across a population of 20 individuals. If 18 of these individuals had a T nucleotide at this site, the frequency of the T allele would be $\frac{18}{20} = 0.9$. The frequency of all alleles at a site sum to *1*. Therefore, if the site is biallelic (*i.e.* has two alleles) then the frequency of the other allele can be calculated as $1 - 0.9 = 0.1$. Due to the ease of working with biallelic single nucleotide polymorphisms (SNPs), and the relative rarity of multiallelic sites, it is common practice to apply the simplifying assumption that all multiallelic SNPs are biallelic, by removing the least common allele(s) (Li, 2011). Alleles within a pair of biallelic SNPs are generally referred to as p and q; p often corresponds to the more frequent allele (the "major allele") but this is not always the case.

Summarising allele frequency differences across the genome

Having derived allele frequencies, a range of summarising statistics exist to quantify allelic diversity within populations. This thesis will utilise four (Figure 1.1). Two fundamental statistics, D_{XY} and π_W, were proposed by Nei and Li in 1979 (Nei and Li, 1979). Within-population diversity (π_W, also known as π) refers to the average number of allelic differences, per site, that would be observed between two sequences drawn at random from a population (Figure 1.1a). Absolute between-population diversity (D_{XY}, also named π_{XY}) refers to the average number of allelic differences, per site, that would be observed between two sequences drawn at random from two distinct populations (Figure 1.1b). By relating π_W and D_{XY}, a third measure, F_{ST}, can be derived. F_{ST} is formally referred to as Fixation Index, and was originally conceived in the 1950s as one of several measures for quantifying homozygosity within populations (Wright, 1965). In the context of genomics, F_{ST} can be thought of more intuitively as relative between-population diversity. Like D_{XY}, F_{ST} measures divergence between populations, but values are adjusted relative to π_W. In practice, this makes it easier to compare sites across the genome, even where local levels of diversity are quite different. A final statistic, D, can be derived by subtracting mean π_W from D_{XY} the of the two populations being compared. D provides an estimate of net nucleotide diversity, or the amount of allelic diversity that has accumulated between populations since they diverged. Like F_{ST}, it relates between-population diversity to within-population diversity, meaning that it is a relative measure of divergence.

(a)

$$\pi_w = p_1 q_1$$

Within-population diversity

(b)

$$D_{XY} = \frac{p_1 q_2 + p_2 q_1}{2}$$

Absolute between-population diversity

(c)

$$F_{ST} = \frac{D_{XY} - \overline{\pi_w}}{D_{XY} + \overline{\pi_w}}$$

Relative between-population diversity

(d)

$$D = D_{XY} - \overline{\pi_w}$$

Net nucleotide diversity

Figure 1.1: Calculation of diversity statistics from allele frequencies

Equations for calculating (a) π_W, (b) D_{XY}, (c) F_{ST}, and (d) Nei's D from allele frequencies p and q.

Challenges of genome scans

The application of diversity statistics to study genome-wide divergence can be generically referred to as a genome scan. Genome scans, comparing pairs of taxa, have generally used F_{ST} as the preferred measure of divergence (Seehausen *et al.*, 2014). The sensitivity of F_{ST} to π_W has a normalising effect on divergence landscapes, making locally elevated "islands" perceivable against a "sea" of low baseline divergence. This ease of detection comes with the cost of conflating D_{XY} and π_W, making it difficult to determine whether peaks are due to within- or between- population diversity. A 2014 paper published by Cruickshank and Hahn demonstrates the pitfalls of drawing conclusions about gene flow on the basis of F_{ST} alone (Cruickshank and Hahn, 2014). Here, they reanalyse published datasets with which F_{ST} comparisons had been used to identify novel islands of divergence. They show that previously identified regions of elevated F_{ST} do not show elevation of D_{XY}, meaning that the observed F_{ST} peaks correspond to low π_W. Fluctuations in π_W and D_{XY} tell very different stories about the evolutionary history of divergence.

Variation in π_W is linked to a range of population history factors. For example, if a population has recovered from an historical bottleneck or founder effect, it may show reduced π_W across the whole genome. Alternatively, low π_W may reflect a selective sweep, where an allele conferring an adaptive advantage has rapidly 'swept' to fixation within a population (Stephan, 2019).

In principle, it would be easier to carry out genome scans based on D_{XY}. Here, elevated D_{XY} would reflect an absolute increase in allelic differences between populations, but not within them. Because D_{XY} accumulates in isolation, and reduces under gene flow, regions showing locally elevated D_{XY} are likely to contain barriers to gene flow. However, whilst D_{XY} is not sensitive to the current π_W, it is sensitive to variation in π_W that arose prior to the most recent common ancestor of the populations being compared. This ancient π_W variation is inherently hard to characterise, being the result of events that occurred too long ago to reliably reconstruct (Cruickshank and Hahn, 2014). This D_{XY} "noise" may obscure genuine signatures of divergence.

A separate, but related, challenge in characterising patterns of genomic diversity is the distribution of intrinsic genomic factors (Wolf and Ellegren, 2017, Foote, 2018). Recombination is the process by which coinherited combinations of alleles (haplotypes) from each parent are rearranged during meiosis. Because recombination creates new combinations of existing variation, it increases the genetic diversity within populations. However, the frequency with which recombination takes place shows significant variation across genomes. In some cases, this is predictable. Extensive studies in *Drosophila* have demonstrated that centromeric, and pericentromeric, regions show a reduced rate of recombination (Begun and Aquadro, 1991, Jensen *et al.*, 2002). This has also been observed in plants such as tomato (Fuentes *et al.*, 2022) and maize (Tenaillon *et al.,* 2002). Recombination rate can affect local levels of genomic diversity by influencing the extent of genomic 'hitchhiking' processes, where evolutionary forces acting on a specific locus within a haplotype block are reflected across the whole haplotype. Studies in a range of organisms have also demonstrated a positive correlation between local mutation rate and π_W (Takahashi, Liu, and Saitou, 2004, McGaugh *et al.*, 2012, Ponnikas *et al.*, 2022).

Even if genomic islands of divergence can be detected, it may not be possible to identify the underlying functional variation (Jiggins and Martin, 2017). Many biological traits are known to be highly polygenic, arising from the activity of large numbers of genes. Even if a conventional genome scan was able to detect all genes involved in a hypothetical polygenic trait, the associated genomic divergence landscape would be too noisy for an experimentalist to interpret at the genetic level. The individual effects of genes may also be so weak that signals are not perceivable. Interpretation may be easier in light of prior knowledge, but this highlights another common issue with genome scans - prior knowledge of the genetic basis of the trait being analysed may lead to biases in interpretation.

Interpreting genomic divergence using trees

I have discussed some of the challenges of characterising genomic diversity, and identifying potential barriers to gene flow that might reveal barrier genes. Because D_{XY} can inform about genetic barriers directly, it represents the most promising genome scan statistic for identifying barrier genes. However, to interpret D_{XY}, a method is needed which reduces the amount of noise arising from historical π_W and differences in intrinsic genomic factors. In describing the issues faced when using relative measures such as F_{ST}, I have demonstrated why simply relating local levels of D_{XY} to π_W are likely to be insufficient to identify barriers to gene flow without additional characterisation. A different approach to interpreting D_{XY} is to compare D_{XY} landscapes across multiple populations. Distinct populations with the same genetic barriers are expected to show shared patterns of D_{XY}, reflecting the genomic regions harbouring barrier genes. By relating D_{XY} landscapes between populations, it may be possible to detect parallel divergence through shared D_{XY} islands, whilst ‘cancelling out’ noise due to ancient, unshared π_W.

Relating genomic divergence across sets of populations has long been carried out through use of dendrograms, more informally referred to as trees. Trees represent the relationship between taxa (which can be individuals or populations) on the basis of pairwise distances between them. In biology, trees typically summarise relationships between taxa by comparing nucleotide or amino acid sequences. Approaches range from simple distance-based measures (such as D_{XY} and D) to sophisticated frameworks modelling evolutionary expectations (such as Maximum Likelihood and Coalescence) (Yang and Rannala, 2012). The increasing availability of whole genome sequencing has seen a shift towards whole genome phylogenetics, where taxa are resolved by comparing as many genomic sites as can be resolved. Whole genome phylogenies are powerful tools for resolving the relationships between populations that are not undergoing gene flow. However, they cannot adequately reflect the mosaic nature of genomic divergence that is expected where gene flow is uneven. For example, consider a tree of four populations showing two distinct ecological specialisations. If these two specialisations arose through a single ancestral divergence event, then phylogenetic analysis of the underlying genes should group the alike populations into monophyletic clades. However, if the populations have since undergone gene flow, then phylogenetic analysis of non-adaptive regions may yield non-monophyletic groupings (Figure 1.2). This phenomenon is broadly termed phylogenetic incongruence (Rokas *et al.*, 2003).

In a 2013 study, Martin *et al.* investigated the patterns of divergence across 31 genomes from sympatric and allopatric *Heliconius* species (Martin *et al.*, 2013). By dividing the genome into 100 kb windows, phylogenetic trees could be constructed across the whole genome, making it possible to predict which regions had been subject to gene flow. By classifying all genomic trees to four predefined topological groups, it was revealed that up to 40 % of trees from sympatric populations showed a topology that was consistent with the geographical distribution of the populations, rather than the whole genome phylogeny, suggesting extensive gene flow. This approach, later dubbed *Twisst* (Martin and Van Belleghem, 2017), has been used in studying a range of divergent traits (for example, Van Belleghem *et* *al.*, 2017, York *et al.*, 2018, Dixon, Kitano, and Kirkpatrick, 2019). *Twisst* summarises trees based on user specified taxa, rather than individual species. For example, in the 2013 study discussed above, 31 individual butterflies (across two experiments) could be simplified into four taxon trees. The numbers of four taxon trees corresponding to different pre-defined hypotheses could then be explored. Because *Twisst* can report multiple topologies, it is robust to incomplete lineage sorting. This process, where organisms unexpectedly group with more distantly related taxa due to shared ancestral variation, has long confounded the study of historical speciation (Sousa and Hey, 2013). Another approach to exploring how phylogenetic trees vary across the genome was developed by Zamani *et al.* (2013) in the *Saguaro* software. This not only summarises tree topologies across the whole genome, but uses machine learning to infer the points at which genome tree topologies change. These approaches differ in their utility. *Twisst* is useful for studying large numbers of individuals based on pre-defined species or geographical relationships. *Saguaro* requires no *a priori* hypotheses about tree topologies, meaning that it can be used to infer genomic landscapes of divergence between organisms without being biased by prior observations. In constructing trees, *Twisst* uses Maximum Likelihood inference in constructing trees. *Saguaro* uses a distance matrix approach, where the genome is summarised into “cacti” that describe the relationship between taxa across consecutive genomic sites. Relating these measures to allele frequencies, as previously described, is challenging. Maximum Likelihood infers tree topologies based on statistical models of sequence evolution, and *Saguaro*’s cacti are not designed to provide any immediate biological meaning. Therefore, drawing inspiration from these approaches, I have developed a genome-wide approach for classifying trees constructed from pooled sequence data, based on D_{XY}.

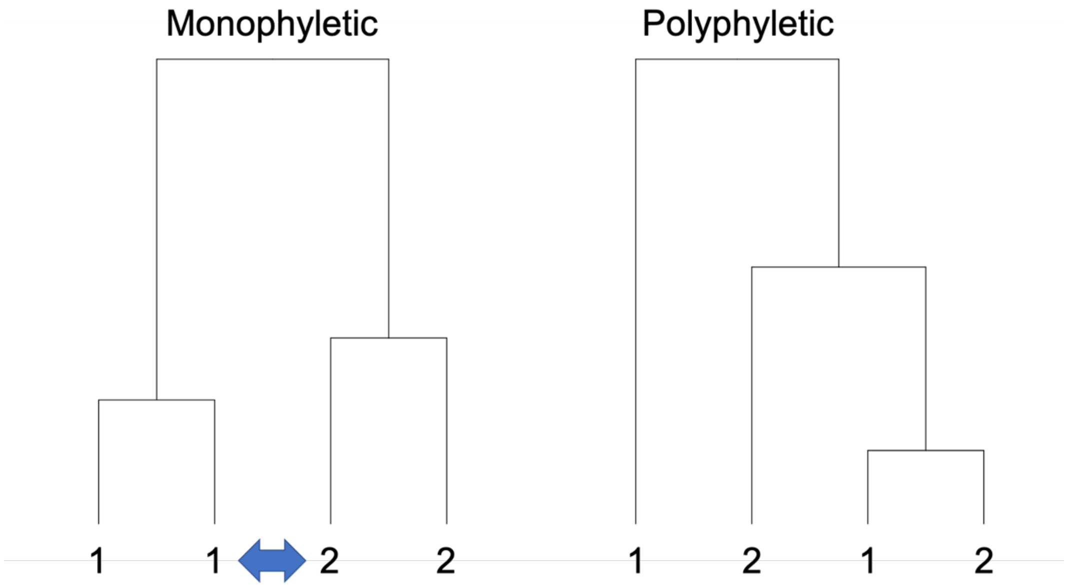


Figure 1.2: Phylogenetic incongruence through gene flow

A hypothetical scenario by which gene flow can result in phylogenetic incongruence. These trees represent whole genome phylogenies constructed for populations showing distinct multigenic traits 1 and 2. Populations showing traits 1 and 2 arose through a single divergence event, resulting in a monophyletic tree (left). However, gene flow (represented by a blue arrow) between populations with different traits reduces divergence around genomic loci that are not involved in maintaining the distinctive identities. The result is a polyphyletic whole genome phylogeny (right). In principle, phylogenies constructed using the genes involved in traits 1 and 2 will still be monophyletic.

The grouping tree scan as a means of identifying barrier genes

To explore the genomic landscape of D_{XY} variation across the genome, I have developed the grouping-tree-scan methodology. This approach summarises D_{XY} across populations and across the genome using UPGMA hierarchical clustering trees. In doing so, the grouping-tree-scan combines the genome-wide nature of a genome scan, with the power of phylogenetic comparisons of multiple taxa. The approach proceeds as follows:

1. Short read DNA is collected from experimental populations, and DNA from each population is sequenced within a pool (pool-seq)
2. Pool-seq data from experimental populations is mapped to a common reference genome

3. Pairwise D_{XY} is calculated between all populations, across overlapping windows of genomic sequence (subgenomic windows)
4. For each window, hierarchical clustering (UPGMA) is used to construct a representative tree
5. Trees are grouped into 'forests', based on topological similarity
6. Forests are summarised based on how their trees group the experimental taxa, and the amount of D_{XY} separating population groups

By analysing multiple populations at once, grouping-tree-scan provides greater power to identify subgenomic regions showing elevated D_{XY}. A focus on pool-seq data means that allelic variation can be sampled and compared across representative samples of populations. Also, population grouping derived from grouping tree scans are 'aphenotypic', meaning that they group populations independently of any phenotypic assumptions.

Antirrhinum as a model system

I have introduced the grouping-tree-scan as a means of detecting parallel signatures of D_{XY} that may reflect genetic barrier loci. I have also characterised two distinct types of genetic barrier: the intrinsic epistatic barrier and the differentially adaptive barrier, that may underpin reproductive isolation within natural populations. By applying the grouping-tree-scan to a suitable model system, it may be possible to test these proposed genetic barrier hypotheses. Analyses detailed here will investigate populations of the garden snapdragon, *Antirrhinum majus*, and its wild relatives. *A. majus*, is an established model system that has been used extensively in studies of classical and molecular evolution (reviewed in Schwarz-Sommer, Davies, and Hudson, 2003). Along with a wealth of classical genetics resources, *A. majus* boasts an ongoing sequencing programme, with a high-quality chromosome level reference genome (Li *et al.*, 2019). *A. majus* is a diploid (2*n* = 16) with a genome of around 500 Mb. Its wild relatives are distributed around the Mediterranean, being primarily native to the Iberian Peninsula. Molecular evidence suggests that the genus *Antirrhinum* did not undergo speciation until relatively recently in evolutionary time, with estimates based on homologous genes in monocots suggesting that this occurred less than 5 million years ago (Vargas *et al.*, 2009). Since then, taxa within *Antirrhinum* have evolved to show a great deal of variation in morphology and growth habit. Most *Antirrhinum* species are self-incompatible, with experimental evidence suggesting that this system is gametophytic and controlled by a polymorphic *S*-locus (Qiao *et al.*, 2004). However, all species are able to interbreed to generate fertile hybrid progeny in the laboratory, and hybrids are also found in natural populations. This combination of recent species divergence, widespread self-incompatibility, and frequent interspecific hybridisation has long confounded taxonomic classification (Wilson and Hudson, 2011). Investigating the genetic barriers reflecting historical divergence events in *Antirrhinum* represents the objective of this thesis.

Detection of genetic barriers involving small RNA loci

In introducing the problem of identifying barrier genes, I have focussed specifically on the detection of allelic variation. However, at least one locus that has been implicated in *Antirrhinum* population divergence is an sRNA locus, which differs in its mode of action compared to protein coding genes. In characterising the molecular basis of yellow pigment biosynthesis in *Antirrhinum*, Bradley *et al.* (2017) demonstrated that floral patterning of yellow was regulated by a small RNA (sRNA) locus, *SULF*. *SULF* restricts the spread of yellow on the face of the flower by facilitating the degradation of the mRNA transcripts of its target gene, *FLA* (previously known as *Am4'CGT*) (Bradley *et al.*, manuscript in preparation). By carrying out cline analysis, which observes changes in allele frequencies across a natural hybrid zone for flower colour, it was demonstrated that the *SULF* genomic region shows steep clines, consistent with natural selection. However, the functional *SULF* element, a 1.5 kb inverted repeat (IR), is absent within yellow-flowered populations. Therefore, differential *SULF* function between populations must not be due to allelic variation, but to the presence or absence of a functional *SULF* IR. If similar loci to *SULF* are involved in genetic barriers, they may not be detected using a grouping tree scan. In Chapter 5, I propose and test a bioinformatic pipeline for identifying *SULF*-like loci through comparison of genome assemblies.

Aims of this thesis

Within the following chapters, I will detail my work using bioinformatic methods to study barriers to gene flow in *Antirrhinum* populations. Using the grouping tree scan approach, I will first test the hypothesis that divergence between sympatric populations of *Antirrhinum majus pseudomajus* and *Antirrhinum majus striatum* is underpinned by an intrinsic epistatic barrier involving differences in flower colour.

In Chapter 4, using an adapted version of the grouping tree scan, I will analyse populations of distantly related *Antirrhinum* species showing distinctive growth phenotypes. I will identify barriers to gene flow between these species and, in doing so, test the hypotheses that populations have diverged through differentially adaptive barriers.

In Chapter 5, I will outline a bioinformatics pipeline that has been developed to test whether *SULF*-like genetic elements are common within *Antirrhinum* genomes, and whether they might be involved in genetic barriers between species.

In light of these observations, the intrinsic epistatic and differentially adaptive barrier hypotheses, and the methods developed to test them, will be evaluated.

2: Materials and methods

DNA extraction and sequencing from pooled leaf tissue

For Whole Genome Sequencing, genomic DNA was isolated using a cetyltrimethylammonium bromide (CTAB) method on ~ 2-5g of leaves harvested either from a single individual or as final weight for pooled samples as described by Coen, Carpenter, and Martin (1986). Samples from the greenhouse were stored at -80°C. Samples collected in field locations throughout France and Spain were either placed in bags in silica, or stored in moist paper towel and kept cool at 4°C until they could be Courier Posted by overnight delivery to the lab in the UK and frozen at -80°C on arrival. Short read sequencing of DNA extracted from pooled leaf tissue was carried out by Novogene using an Illumina HiSeq 2500. A GPS reading was taken at each wild population sampling location. These are recorded in Appendix 1.

Reference genome assembly

Analyses in Chapter 4 and Chapter 5 used the published *Antirrhinum majus* reference genome (Li *et al.*, 2019). Analyses in Chapter 3 used a more recent draft of the *Antirrhinum majus* reference genome, which I have termed *A. majus* Reference genome version 3.5 (V3.5). This draft was assembled by Sihui Zhu in the group of Yongbiao Xue at the Beijing Institute of Genomics. *Antirrhinum majus* JI7 was grown, and leaf tissue harvested, as specified within the Methods section of the published *A. majus* reference genome paper (Li *et al.*, 2019). DNA was extracted from leaf tissue using the CTAB method, and sequenced. 40X PacBio HiFi reads were generated, and assembled using FALCON (Chin *et al.*, 2016). FALCON-Unzip was used for initial assembly of reads. Hi-C data was generated by Novogene, and used with FALCON-phase to determine the phase of all contigs. Phased contigs were then analysed using optical Bionano molecular maps. Low quality maps, with length ≤ 150 kb or label number ≤ 9 were removed. Bionano genome maps, combined with two phased haplotype assemblies, were passed to the Bionano Solve hybrid scaffolding pipeline (version 3.6). This was run in non-haplotype-aware mode, as recommended within the manual. Where conflicts were detected between sequence maps and optical molecular maps, both were cut at the conflict site and assembled, with parameters `-B2` and `-N2`.

Adapters were removed from raw Hi-C reads, and low-quality bases were trimmed using Trim Galore! (version 0.6.1) (Krueger, 2015) with default parameters. Clean reads were mapped to the Bionano assembly using BWA-MEM (Li and Durbin, 2009). To generate a chromosomal assembly, the 3D-DNA pipeline (Dudchenko *et al.*, 2017) was first used to refine the assembly. Manual review of the candidate assembly was carried out interactively using Juicebox Assembly Tools (Durand *et al.*, 2016). The reviewed chromosomal assembly was then generated using the following command: `run-asm-pipeline-post-review.sh -s finalize --sort-output --build-gapped-map`

Mapping reads to reference genome

Pooled Illumina sequencing reads in FASTQ format were mapped to the reference genome using BWA-MEM with the `-M` flag for Picard (http://broadinstitute.github.io/picard/) compatibility. Mapped SAM files were sorted using SAMtools, and duplicate reads were removed using Picard MarkDuplicates. Local realignment around indels was carried out using GATK (McKenna *et al.*, 2010) RealignerTargetCreater to generate an intervals file and GATK IndelRealigner to carry out the realignment. Read coverage was determined for each pool using GATK DepthOfCoverage. Scripts used to run the mapping tools are included in the accompanying shell script, *snap_map.sh* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

SlidingWindows analysis

MPILEUP files were generated from processed BAM files using SAMtools (Danecek *et al.*, 2021) mpileup, with minimum and maximum quality thresholds of 30 and 40 respectively. The `-B` flag was used to disable the use of probabilistic realignment in the computation of base alignment quality, as this can result in an increase in false SNP calls due to misalignments. The `-A` flag was also set to include orphaned reads in variant calling. MPILEUP files were converted to Popoolation2 SYNC format (Kofler, Vinay Pandey, and Schlötterer, 2011) for compatibility with SlidingWindows.py (version 1.10), a population genomics Python script developed by David Field. SlidingWindows.py is available on GitHub (https://github.com/dfield007/slidingWindows). SlidingWindows calculates populations genetic statistics across the genome, in user defined window sizes. Unless otherwise specified, analyses conducted here were carried out using window sizes of 50 kb, with a 25 kb overlap. Minimum and maximum quality thresholds were set at 20 and 400 respectively. For an allele to be called at a site, it must be supported by at least two reads in at least two populations. Within these analyses, I used three statistics calculated by SlidingWindows; “piAdj” is π_W, adjusted for binomial sampling. piAdj = $\pi_W[\frac{(M-1)}{(M-2b+1)}]$, where M is the read depth, and b is the minimum read count required for an allele to be called. “dXYraw” is D_{XY}. “FstfromMeanPiAdj” is F_{ST} calculated using adjusted π_W values. D is calculated by subtracting mean π_W from D_{XY}. Example code used to run SlidingWindows is included in the accompanying shell script, *sync_SlidingWindows.sh* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Constructing a mean D_{XY} / D tree

Scripts to process the output data tables from SlidingWindows were written in R (version 4.1.3), using the RStudio IDE (version 1.4.1717). Between population statistics were used to populate distance matrices corresponding to each genomic window (“dXYraw”, and “FstfromMeanPiAdj” were used; see the README at https://github.com/dfield007/slidingWindows). For a given distance matrix, an UPGMA hierarchical clustering tree was constructed using the agnes function from the cluster package (Maechler *et al.*, 2022). To generate a mean tree for multiple genomic windows, or the whole genome, the mean was calculated across distance matrices using the Reduce function. Example code for generating mean trees is included within the accompanying R script, *grouping_tree_scan.R* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Whole genome sweep analysis

To simulate the effect of a whole genome selective sweep within a population, the SYNC file for each chromosome was, in turn, parsed to extract the column corresponding to the MP11 population. The allele frequencies across each of these columns were then edited using a Python script. For each genomic site where more than one allele showed depth > 0, one allele was randomly sampled (frequencies of N / del alleles were ignored). Probability of sampling was weighted based on the frequency of alleles present. The depth of the sampled allele was set to equal the total depth at that site. The depth of all other alleles was set to 0. Each edited table was then used to replace the original columns within the SYNC files, and the SlidingWindows analysis run again. Code used to carry out the whole genome sweep is included in the accompanying python program, *artificial_sweep.py* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Grouping of forests based on comparisons to seed trees

The similarity of UPGMA trees was estimated using the cophenetic correlation coefficient (Sokal and Rohlf, 1962). To compare two trees, their cophenetic matrices were obtained using the cophenetic function in base R. Cophenetic matrices were then compared based on their Pearson correlation coefficient, $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$, where x and y correspond to the minimum merging distances in each matrix. All subgenomic trees were clustered into forests by iterative comparisons to randomly sampled seed trees. To do this, a tree was sampled at random, and compared to all other subgenomic trees using the cophenetic correlation coefficient. An $r$value of 0.5 was chosen as the threshold for tree similarity, in order to capture trees showing moderate or high topological similarity. Where r> 0.5, trees were declared similar, and grouped. Once grouped, trees were unable to participate in subsequent comparisons. Once a seed tree had been compared to all genomic trees, it was added to the forest, and another seed tree was selected. This was repeated iteratively until no trees remained. Example code for clustering trees into forests is included within the accompanying R script, *grouping_tree_scan.R* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Calculation of shortest root branch (SRB)

To calculate the length of the shortest root branch (SRB) for a tree, a cophenetic matrix was derived using the base R cophenetic function. SRB is equal to the maximum value within the cophenetic matrix (the tree height) minus the second highest value. Example code for calculating SRB is included within the accompanying R script, *grouping_tree_scan.R* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Bootstrapping forest clustering

To test how replicable the results of the forest clustering are, a simple bootstrapping process was implemented. Forest clustering was run 25 times. Each time, a mean SRB value was calculated for each forest, by averaging the SRB values of all trees within a given forest. The forest showing the highest mean SRB was identified, and all subgenomic regions within it recorded. This was repeated 24 more times, after which all subgenomic regions that have appeared within the most outlying SRB forest were reported. Example code for bootstrapping the forest clustering process is included within the accompanying R script, *grouping_tree_scan.R* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Classifying trees based on root division

Trees were divided at their topmost branch (the root branch) using the cutree function, with k (the desired number of groups) set to two. To classify trees based on root division, populations within each of the split groups were recorded, and compared to a user specified signature. Example code for carrying out root division classification is included within the accompanying R script, *grouping_tree_scan.R* (https://github.com/DR-Antirrhinum/DR_thesis_2023).

Growth conditions and crosses

Plant populations used in these studies were managed by Lucy Copsey, Desmond Bradley, and the John Innes Centre Horticultural Services department. Greenhouse plants were grown in JI compost-soil mixes as described (Carpenter *et al.*, 1987) with supplemental lights in winter to give 12-16 hour days. Outside plants were grown in summer on the same soil mixes in pots or plugs in trays on raised benches. Lines from self-incompatible species were maintained by inter-sibling crossing. For crosses, young floral buds (pre-anthesis) were opened with sharp forceps and young anthers removed. Two to four days later when the flower was open, pollen from another plant was daubed onto the stigma using forceps carrying pollen or the whole stamen. Four to six weeks later ripe capsules with mature seed could be harvested.

KASP genotyping

KASP Genotyping was performed by Desmond Bradley, as described in Bradley *et al.* (2017). The fluorescence signals discriminating the two alleles were detected in a BioRad CFX96 light cycler and data processed with the BioRad CFX Manager software v3.1. AFLP methods used standard PCR with a PCR cycle 1 at 94°C for 3 minutes, followed by 35 cycles of 94°C for 1 minute, 55-58°C for 1 minute and 72°C for 2 minutes, before a final cycle of 72°C for 10 minutes and storage at 16°C until collection. PCR products were run on 1 % (*w*/*v*) agarose gels and stained with ethidium bromide before standard UV imaging. KASP / AFLP oligos are recorded in Appendix 3.

Flower photography

Flower photography was carried out by Desmond Bradley. Flowers were places on black velvet with a scale bar and Small Grey & Colour Separation Chart (Danes Picta BST13) for colour, light level and white balance monitoring. Desmond used an Olympus XZ-1 (10 Megapixels) Camera with side / overhead lighting via table lamps fitted with halogen 42W 630 lumen (2800k) warm white light bulbs. Camera settings were set to the closest White Balance of 3000K, no flash, Macro On, F stop 8.0, Exposure Time 1/20-1/40 sec, ISO 200, RAW images, aspect 4:3, high definition.

RNA extraction from flower buds

Desmond Bradley isolated total RNA from various tissues using the RNeasy Plant mini Kit (Qiagen). Leaves were harvested from 3-10 individual plants in a range of leaf sizes from very small to mature (Total 10-20 leaves / individual). Shoots (lateral shoot branches from leaf axils at various stages of growth) where harvested from 3-10 plants, collecting the shoot tips (~ top 1 cm) (Total 20-40 / individual). Flower buds (including all floral organs) in a range of developmental sizes from small to just open flower were harvested from 3-10 plants (Total 10-20 buds / individual). Mature flowers from just before opening to just opened, from 3-10 plants, were separated into tubes from lobes. Lobes were then either pooled from 3 different individuals of the same genotype, or further separated into dorsal/upper lobes vs. ventral/lower lobes before pooling. (Total 2-20 /individual). Three independent replicates were collected for all genotypes except in the analysis of *A. m. pseudomajus, A. m. striatum*, and *A. molle*, where a single sample was used to ascertain general sRNA composition. For *A. m. pseudomajus* compared to *A. m. striatum* Desmond used A*. m. pseudomajus* Accession Ac1266 (sub-location M-AUT-8) and *A. m. striatum* Accession Ac1125 (sub-location Z-ALE-11). Total lobes were harvested from various individuals to give pools of each genotype in triplicate.

Small RNA libraries were constructed by Desmond Bradley as described in Bradley *et al.*, 2017. For a comparison of single samples of *A. m. pseudomajus* Ac1099 (sublocation Z-NDM-4), *A. m. striatum* Ac1130 (sublocation Z-VCO-4) and *A. molle* Ac1313 (sublocation W-RGA-34). Desmond used pools of 3-10 plants in each case, and dissected out petals from a range of bud sizes from very small to just opened flowers (10-20 buds / individual). For a detailed study of *A. m. pseudomajus* Ac1099 (sublocation Z-NDM-4) versus *A. sempervirens* Ac1169 (sublocation C-NAP-364) various tissues were harvested in triplicate, each from 3-10 individual plants. This gave 10-40 tissue samples / individual. Leaves were harvested in a range of leaf sizes from very small to mature. Floral buds were harvested in a range of total floral buds from smallest buds to first open flower. Petals were harvested and tubes discarded to collect dorsal/upper half lobes and ventral/lower lobes from a range of floral buds from smallest buds to first open flower. Shoots here harvested at their tips (~ top 1 cm) from various sized branches in the axils of leaves. Sequencing of sRNAs was carried out by Maria-Elena Mannarelli at the University of East Anglia.

RNAseq differential expression analysis using DESeq2

Differential expression analysis of total RNA between *A. m. pseudomajus* and *A. m. striatum* was carried out by Annabel Whibley. DESeq2 (Love, Huber, and Anders, 2014) was run using default parameters.

Gene prediction using eggNOG-mapper

To automatically derive functional predictions, eggNOG-mapper 2.1.9 (Cantalapiedra *et al.*, 2021) was used, through the web interface (http://eggnog-mapper.embl.de). Sequences to be annotated were submitted in FASTA format, selecting the genomic data input option. Blastx-like was chosen as the gene prediction method. Otherwise, default parameters were used.

Eight species genome assemblies

Genome assemblies of the eight species analysed in Chapter 5 were assembled by Sihui Zhu at the Beijing Institute of Genomics. Illumina reads were sequenced, and trimmed using Trim Galore! with parameters `-q25 --stringency 3` to remove low-quality bases and adapters. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to evaluate species reads based on the k-mer spectrum, heterozygosity rate and haploid genome length. Contig-level assemblies were performed using PacBio reads and the Canu package (version 1.9) (Koren *et al.*, 2017). purge_dups (Guan *et al.*, 2020) was used to remove duplicate contigs in primary assemblies. Repetitive sequence within the assemblies was identified and masked using RepeatMasker (Smit *et al.*, 2013). High-quality plant proteins were retrieved from SwissProt and aligned to the genome using ProtHint (Brůna *et al*., 2020). RNAseq datasets were also aligned to the repeat masked genomes. Gene prediction was carried out using the Braker2 (Brůna *et al.*, 2021). The overall quality of genome assemblies and genome annotations were assessed using LAI (Ou *et al.*, 2018) and BUSCO (Simão *et al.*, 2015).

Detection of genomic inverted repeats

IRs were identified within genome assemblies using Inverted Repeat Finder (IRF) (Warburton *et al.*, 2004), with the following parameters: `irf307.dos.exe in.fa 2 3 4 70 10 150 30000 3500 -d -a4 -i2`

Trimming and mapping sRNAs

sRNAs were first trimmed of High Definition adapters (Xu *et al.*, 2015) using Trim Galore!. To do this, the first seven nucleotides of the Illumina adapter sequence were matched (TGGAATT), along with the four HD nucleotides at the 3' and 5' ends of the inserted sequence. Trimmed sRNAs were mapped to all IRs using Bowtie, allowing no more than two mismatches between the sRNA and the IR (`-v 2`).

BLASTN searches

Local BLASTN searches were carried out using the BLAST+ application (Camacho *et al.*, 2009), using blastn 2.9.0+. Predicted protein coding sequences for *Antirrhinum majus* were downloaded from the Antirrhinum genome database website (http://bioinfo.sibs.ac.cn/Am/index.php).

Mapping IRs to genome assemblies

IRs were mapped to all species genome assemblies, and the *A. majus* reference genome, using Minimap2 (Li, 2021), with the following parameters: `./minimap2 -x asm5 -a -k 15 -w 5 target.fa query.fa`

BLASTX searches

Comparisons against the NCBI non-redundant protein database (nr) were carried out using the online BLASTX search tool (https://blast.ncbi.nlm.nih.gov/).

3: Grouping tree scans reveal barriers to gene flow between two *Antirrhinum* subspecies

Introduction

Natural hybrid zones, where distinct populations meet and undergo gene exchange, have long been a valuable resource in the study of speciation. Over time, pervasive gene flow can have an extensive homogenising effect on population genomes. Against this backdrop of reduced divergence, loci which resist gene flow show characteristic islands of elevated relative divergence. By carrying out comparative genomic analyses of populations either side of a hybrid zone, it is theoretically possible to identify barriers to gene flow *in silico*, and thereby capture the mechanistic basis of the population divergence (Wolf and Ellegren, 2017, Ringbaur *et al.*, 2018). This approach has been utilised in *A. majus*, of which several natural hybrid zones between the contrasting magenta- and yellow-flowered subspecies, *Antirrhinum majus pseudomajus* and *Antirrhinum majus striatum,* have been identified within the north of Spain. *A. m. pseudomajus* has magenta flowers with yellow highlights at the point of bee entry, whereas A. *m. striatum* has yellow flowers with magenta veins around the bee entry point (Figure 3.1). Between the two population is an area of around 1 km where predominantly hybrid colour phenotypes are detected (Whibley *et al.*, 2006).

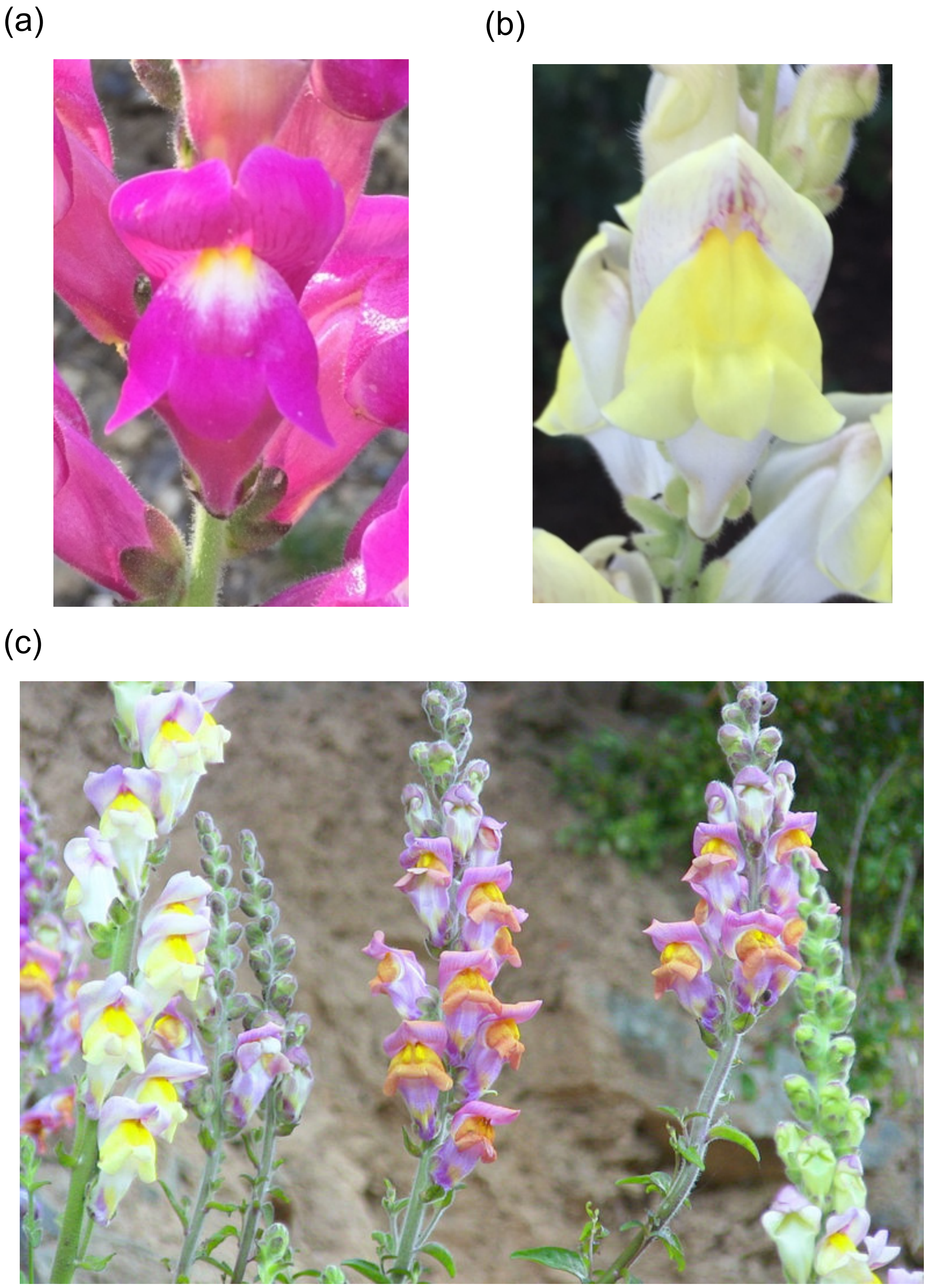


Figure 3.1: Photographs of A. m. pseudomajus and A. m. striatum flowers

Photographs of flowers from (a) *A. m. pseudomajus*, (b) *A. m. striatum*, and (c) hybrid plants.

Flower colour in *A. m. pseudomajus* and *A. m. striatum* makes use of two distinct pigments – magenta coloured anthocyanin, and yellow aurone. The biosynthesis pathways of anthocyanin and aurone have been extensively studied in *Antirrhinum* (Martin *et al.*, 1991, Schwinn *et al.*, 2006, Davies *et al.*, 2006, Nakayama, 2022). A subset of biosynthetic enzymes have been shown to be major regulators of flower colour patterns, and some have been directly implicated in phenotypic divergence in the field (Schwinn *et al.*, 2006, Whibley *et al*, 2006, Bradley *et al.*, 2017, Tavares *et al.*, 2018). A major regulator of magenta pigmentation is the *ROSEA* (*ROS*) locus. This region contains the *MYB-like* genes *ROS1* and *ROS2*, which interact with components of the anthocyanin pathway to generate magenta pigmentation throughout the flower (Figure 3.2a). *VENOSA* (*VE*) also encodes a MYB-like transcription factor. *VE* expression promotes the synthesis of anthocyanin specifically along the veins of dorsal petals. Allelic variation at *ROS* and *VE* has been shown to generate varying intensities of magenta colouration, in a manner that is influenced by environmental conditions (Schwinn *et al.*, 2006). *ELUTA* (*EL*), a MYB-like semidominant repressor of anthocyanin pigmentation, is tightly linked to *ROS*. *EL* has a restricting effect on the spread of magenta by both *ROS* and *VENOSA* (Tavares *et al.*, 2018) (Figure 3.2c). Two loci have been implicated in controlling yellow flower colour. *FLAVIA* (*FLA*) catalyses the biosynthesis of the aurone precursor. *FLA* activity results in spread yellow throughout the flower (Figure 3.2b). Mutant *fla* alleles show restricted or null yellow (Bradley *et al.*, manuscript in preparation). *SULF* is a sRNA producing IR that acts to restrict yellow pigmentation by targeting the transcripts of *FLA* for degradation (Bradley *et al.*, 2017) (Figure 3.2c).

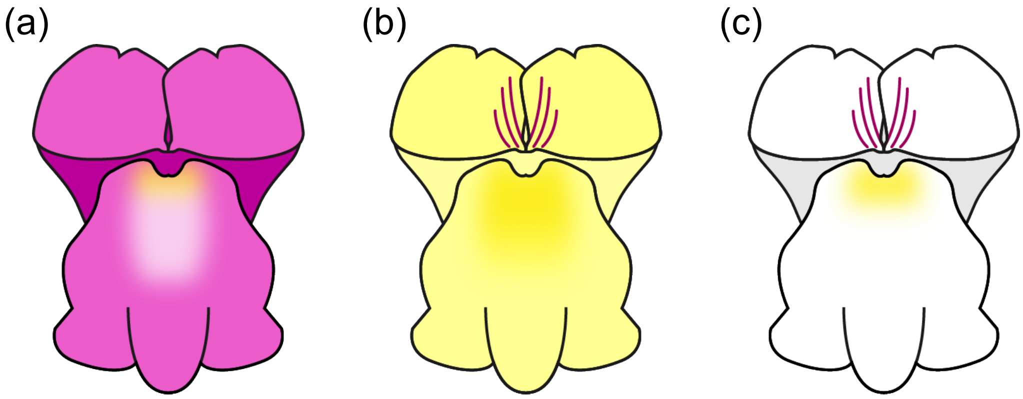


Figure 3.2: Spread of magenta and yellow in Antirrhinum flowers

Illustrations showing (a) spread magenta, (b) spread yellow, and (c) restricted yellow and magenta in the flowers of *A. m. pseudomajus and A. m. striatum.* Flower illustrations by Mabon Elis.

Distinct populations of *Antirrhinum majus* subspecies show equivalent fitness in a shared environment

Within and around the Planoles hybrid zone, a mutualistic relationship exists between snapdragons and several bumblebees of the *Bombus* genus (Tastard *et al.*, 2008). The characteristic enclosed flowers of *Antirrhinum* species prevent pollination by small insects, thereby securing a large nectar reward for bumblebees. This is believed to promote faithful visitation (Vargas *et al.*, 2017), provided that flowers are perceivable to bumblebees. In laboratory experiments, the known pollinator *Bombus terrestris* does not significantly favour *A. m. striatum* or *A. m. pseudomajus* (Jaworski *et al.*, 2015). The extent to which this approach informs about pollinator preference in the field is unclear, as it is infeasible to faithfully reconstitute complex ecological relationships where other factors such as scent might be having cryptic effects on bee preference. Field studies have characterised some of the complexities within the relationships between plants and pollinators, including the prevalence of seed and nectar robbing by predatory weevils and parasitic bees (Leonard *et al.*, 2013, Jaworski, Thébaud, and Chave, 2016) as well as the possible effects of frequency dependent selection, whereby pollinator attraction is rooted not only in colour variation within individuals, but within groups of organisms showing common phenotypes (Tastard *et al.*, 2012). However, no clear evidence exists to suggest that either subspecies is significantly disadvantaged relative to the other (Khimoun *et al.*, 2011). Intensive genomic and ecological study at one of these hybrid zones, near the village of Planoles, has provided evidence that the two subspecies are broadly genomically similar, but diverge at three distinct loci containing four genes involved in flower colour (Bradley *et al.*, 2017, Tavares *et al.*, 2018, Bradley *et al.*, preliminary data). The notion that colour genes alone might have the capacity to initiate and participate in population divergence represents an exciting prospect, informing not only about the diversification of flowering plants, but also about the properties of barrier genes. Other experimental systems, including *Heliconius* butterflies and *Mimulus* monkeyflowers, show strong evidence of divergence around colour loci (Cooley and Willis, 2009, Brien *et al.*, 2022).

The Planoles hybrid zone has been studied using clines and F_{ST} analysis

A classic molecular technique in the study of natural hybrid zones is cline analysis (Barton and Hewitt, 1985). By sampling allele frequency across the span of the hybrid zone, it is possible to detect spatial patterns in allelic distribution. Where allelic variation is strongly involved in phenotypic diversity between populations undergoing hybridisation, this spatial pattern, or cline, will be distinctly steep, reflecting the rapid increase (or decrease) in the frequency of the population-specific allele with geographic position. Cline analyses at the Planoles hybrid zone have shown that steep clines exist at several flower colour genes, suggesting that these loci are under selection (Bradley *et al.*, 2017, Tavares *et al.*, 2018). However, steep clines can also indicate that minimal gene flow has taken place across the hybrid zone, for example if the two subspecies met only recently. Tavares *et al.* (2018) carried out whole genomic divergence comparisons using F_{ST}, a measure of relative sequence divergence. F_{ST} analysis of populations from either side of the Planoles hybrid zone has shown that the genomes of these two distinct subspecies are strikingly homogeneous, except at a small subset of loci which show substantial divergence. By combining F_{ST} comparisons with cline analyses and classical genetics experiments, it has been shown that two of these divergence peaks correspond to *ROS* and *FLA* (Tavares *et al.*, 2018, Bradley *et al.*, manuscript in preparation).

Population analyses can be subject to ascertainment bias

An ongoing challenge of determining speciation genes through analysis of hybrid zones is ascertainment bias, where certain members of a population are more likely to be sampled than others. In biology, ascertainment bias is commonly discussed in terms of detecting SNPs (Lachance and Tishkoff, 2013). In principle, ascertainment bias can manifest when the methodology used to capture candidate loci is designed with the segregating phenotypes in mind. In *Antirrhinum*, the techniques used to detect divergence at colour loci were designed with respect to the flower colour phenotypes of the populations or individuals. Plants were chosen for sampling by observing the colour of their flowers, but not on the basis of any other trait. Variation in other traits, particularly those that were not visible, was ignored. If, for example, divergence was also driven by a trait such as root length or floral scent, and neither of these traits showed the same pattern of variation as flower colour, then this would not necessarily be reflected in the detected divergence islands.

Hypotheses for explaining colour gene divergence

To explain the apparently exceptional nature of colour gene loci, two hypotheses can be tested. The first postulates that divergence at colour gene loci is driven by as-yet-undetected environmental factors. In this case, it is predicted that comparisons between populations of differing geographical origin will reveal barriers at different traits, depending on the local environmental conditions of the populations tested. These additional barriers have not been detected at the hybrid zone due to ascertainment bias. An alternative hypothesis suggests that colour genes underpin a genetic barrier between *A. m. pseudomajus* and *A. m. striatum.* A simple architecture that may give rise to such a barrier is heterozygote disadvantage, where non-parental allele combinations are less fit than parental types (Láruson and Reed, 2016). The dynamic of a barrier causing heterozygote disadvantage will vary depending on the nature of the genes involved. However, the effects on fitness at different loci are expected to be additive. Alternatively, an intrinsic epistatic barrier may be in place. In this case, barrier loci are coadapted, and affect fitness multiplicatively (Csilléry *et al.*, 2018). If the gene flow barrier is manifested through epistasis, it is expected that divergence islands at a panel of colour genes will always be detectable in comparisons between subspecies, regardless of local environmental conditions. Furthermore, assuming that epistatically-maintained gene flow barriers are not a common phenomenon, any newly identified islands will be predicted to correspond to additional colour genes. This is in contrast to the heterozygote disadvantage case, where islands at genes relating to other phenotypes might be expected.

Aim of this work

In this Chapter, I will carry out a grouping tree scan to test the hypothesis that the reproductive barrier between *A. m. pseudomajus* and *A. m. striatum* is underpinned by an intrinsic epistatic barrier involving colour genes. Using pool-seq data from 18 plant populations, I will first observe the extent of divergence between populations by generating a whole genome D_{XY} tree. Then, to detect barriers to gene flow between populations, I will carry out a grouping tree scan. Subgenomic trees will be aphenotypically clustered into forests, based on how they group the populations. Forests containing trees which show elevated between-group divergence will be analysed. This will inform about which genomic regions show elevated divergence. It will also show whether elevated divergence is between *A. m. pseudomajus* and *A. m. striatum*, or between populations of the same subspecies. The genic content of genomic islands of divergence will be analysed. If identified regions contain genes involved in flower colour, this will provide evidence that divergence between populations might be driven by an intrinsic epistatic barrier underpinned by colour genes. It will also demonstrate that prior analyses of flower colour at the Planoles hybrid zone have not been confounded by ascertainment bias. If more loci are detected, or barriers are detected between alike populations, this will demonstrate that cryptic environmental or genetic factors may be involved in differentiating the populations.

Results

Results: D_{XY} tree scans reveal candidate barrier regions

To explore the relationship between nine populations of *A. m. pseudomajus* and nine of *A. m. striatum* (Figure 3.3), whole-genome genetic distance trees were generated. For each population, leaves were sampled from 20-60 individuals and pooled (pool details recorded in Appendix 1). DNA was extracted from each pool, sequenced, and mapped to V3.5 of the *Antirrhinum majus* reference genome (Zhu *et al.*, preliminary work). SNPs were filtered to ensure all were biallelic (for sites with more than two alleles, only the two most common alleles were considered). In comparing two populations, POP1 and POP2, for a given SNP, p_1 and q_1 are the frequencies of each allele in POP1, and p_2 and q_2 as the frequency of each allele in POP2. Genetic distance relationships between a given set of populations can be summarised by constructing a UPGMA tree (Sokal and Michener, 1958).

The primary goal of this analysis was to identify possible barriers to gene flow. To minimise confounding effects of inbreeding and selective sweeps on distance trees, genetic distance has been reported using Nei's D_{XY} (Nei, 1987). D_{XY} is defined as $\frac{p_1q_2+p_2q_1}{2}$ averaged over all sampled positions. Constructing a D_{XY} tree based on the whole genome sequence of each population gave a polyphyletic grouping for the subspecies, indicating no genome-wide barrier to gene flow between subspecies (Figure 3.4).

One disadvantage of using D_{XY} for tree construction is that does not evaluate to 0 when comparing identical populations. For identical populations, $\mathrm{p}_1 = \mathrm{p}_2$, and $q_1 = q_2$ so D_{XY} equates to within population diversity $\pi_w = p_1q_1 = p_2q_2$ averaged over all sampled positions. The non-zero distance between identical or similar populations accounts for the elongated terminal branches in D_{XY} (Figure 3.4). Trees generated using a different distance measure, such as Nei's standard genetic distance $D = D_{XY} - \overline{\pi_w}$ do not suffer from this problem (Figure 3.5). D gave a different tree to that for D_{XY}. Again, subspecies were grouped polyphyletically but in this case tree topology may be influenced by sweeps and / or inbreeding.

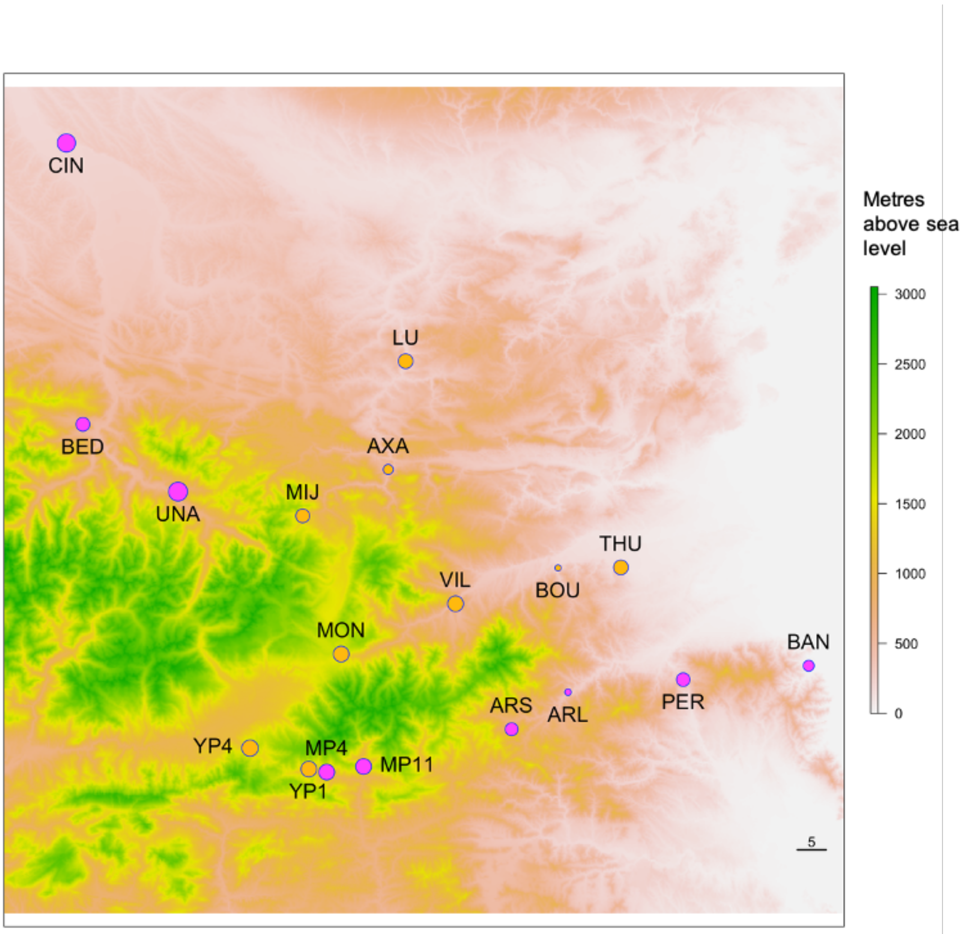


Figure 3.3: Map of 18 experimental populations of A. m. pseudomajus and A. m. striatum

Sampling locations of nine populations of *A. m.* pseudomajus, and nine of *A. m. striatum*. Terrain is coloured according to altitude. Magenta points are *A. m. pseudomajus* populations, yellow points are *A. m. striatum* populations. Points are scaled according to number of sampled individuals (see Appendix 1).

To illustrate the insensitivity of D_{XY} trees to inbreeding / sweeps, one of the populations, MP11, was subjected to a genome-wide sweep by randomly sampling alleles at each position and setting their frequency to 1 (*i.e.* fixing them). The resulting D_{XY} tree showed only minor changes (probably caused by applying the sweep prior to filtering for biallelic sites), whereas in the D tree, MP11 was shifted to become the outgroup (Figure 3.6).

To determine whether all regions in the genome gave similar D_{XY} trees, the genome was divided into 50 kb windows, with a 25 kb overlap between adjacent windows, yielding 19,520 windows in total. Figure 3.7 shows a D_{XY} tree for a randomly sampled window. This tree has a different topology compared to the whole genome tree, but whether this is caused by sampling a smaller region of the genome or differential gene flow is unclear.

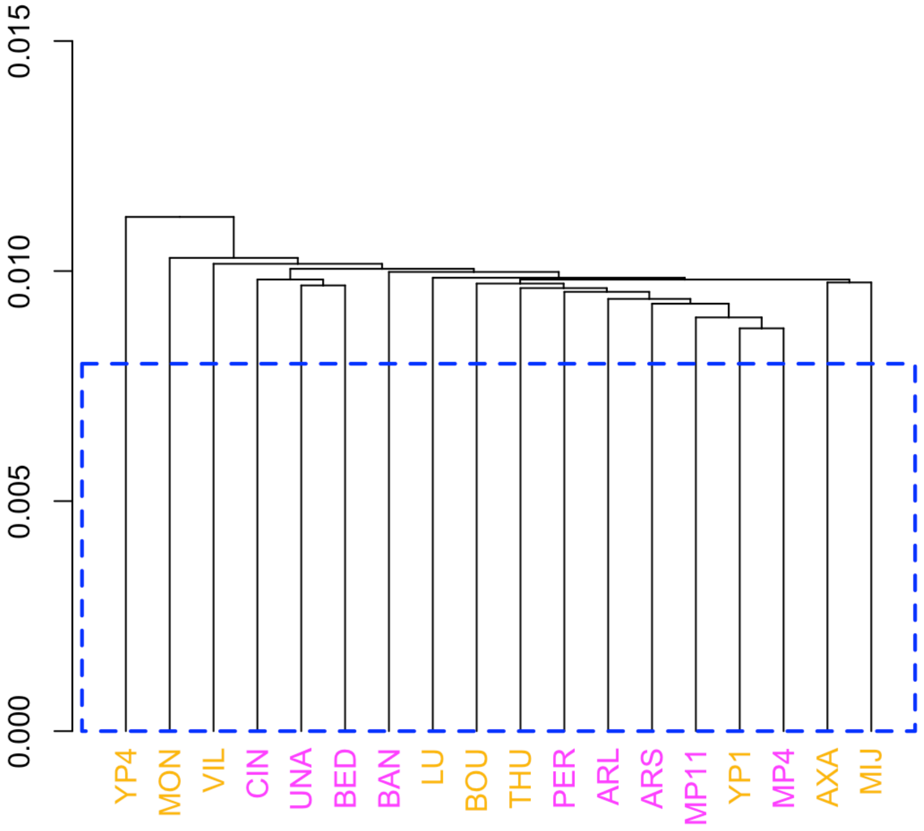


Figure 3.4: Whole genome mean D_{XY} tree for 18 populations

A D_{XY} tree summarising the relationships between the 18 experimental populations. Mean D_{XY} between each pair of populations was averaged over all 19,520 50 kb genomic windows. The blue box indicates the minimum whole genomic mean value of π_w across the 18 populations. Population names in magenta are *A. m. pseudomajus*, names in yellow are *A. m. striatum*.

Genomic regions resistant to gene flow between two groups of populations might be expected to give D_{XY} trees with a deep division between these populations. To identify such trees, all 19,520 D_{XY} trees were first classified according to their cophenetic correlation coefficient, a measure of tree similarity (Sokal and Rohlf, 1962). To avoid having to compute pairwise comparisons of all 19,520 trees, which would take a long time computationally, trees were grouped based on similarity to randomly sampled seed trees. To do this, a random D_{XY} tree was first selected as a seed and compared to all other genomic trees in a pairwise manner. Each comparison involved creating a cophenetic matrix for each tree, each being populated with the minimum merging distances between all population pairs in that tree (Figure 3.8). The cophenetic correlation coefficient between two trees was then calculated from the linear correlation between a pair of cophenetic matrices (Figure 3.9). All trees with a correlation > 0.5 were classified together and removed from further comparisons. Another seed tree was then randomly selected from the remaining unclassified trees. This process was iterated until no trees remained, yielding groups of trees, henceforth termed forests. Running this algorithm until no trees remained yielded 593 forests.

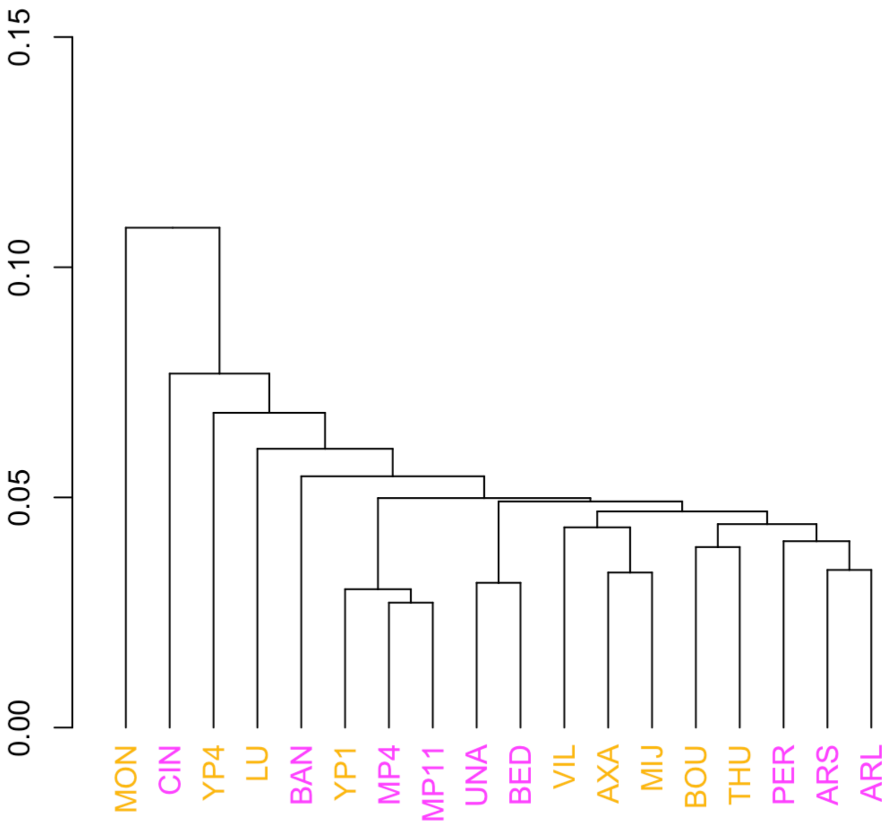


Figure 3.5: Whole genome D tree for 18 populations

A D tree summarising the relationships between the 18 experimental populations. Mean D between each pair of populations was averaged over all 19,520 50 kb genomic windows. Population names in magenta are *A. m. pseudomajus*, names in yellow are *A. m. striatum*.

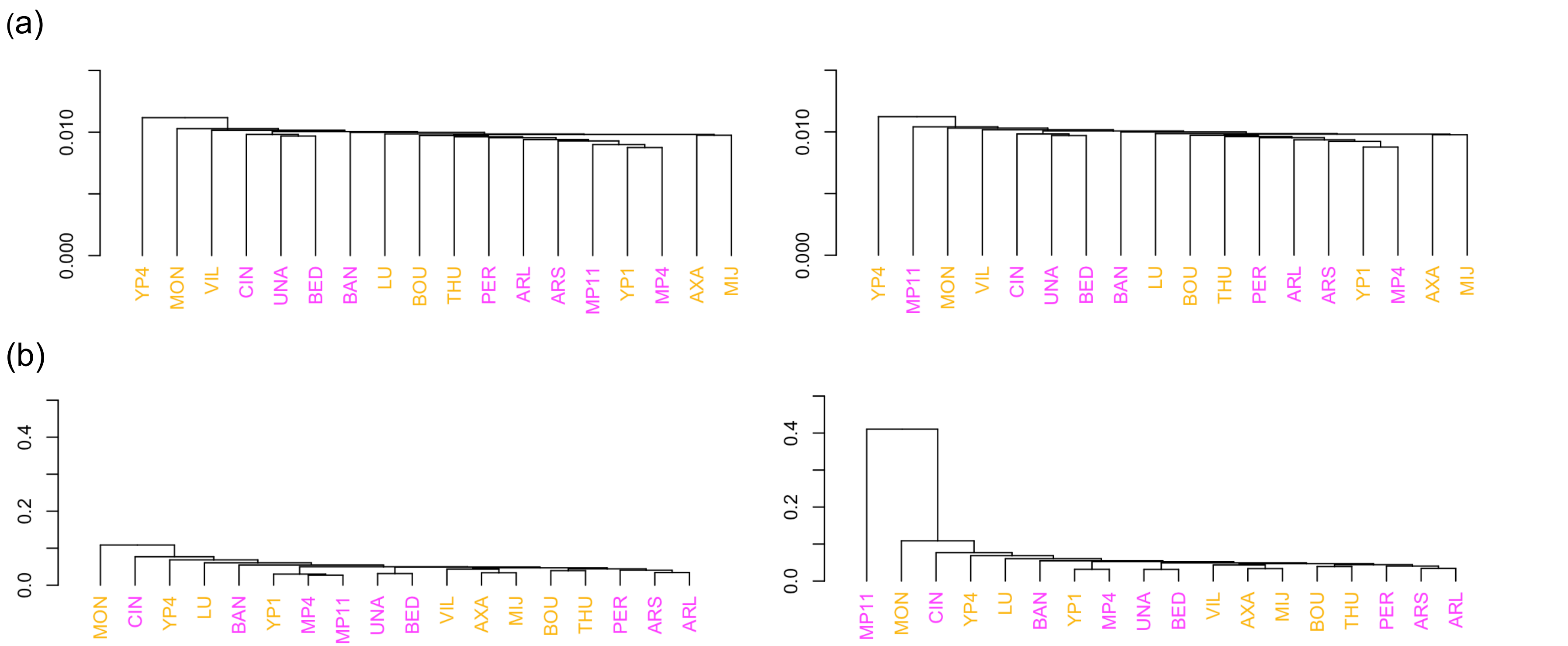


Figure 3.6: Trees plotted before and after a whole genome selective sweep in the MP11 population

Trees plotted before and after a whole genome selective sweep has been applied to the MP11 population. Within the sweep, all polymorphic sites within the MP11 genome are fixed for an allele that has been sampled from the pool of variation at that site. (a) Mean whole genome D_{XY} tree before (left) and after (right) the sweep. (b) Mean whole genome D tree before (left) and after (right) the sweep. Population names in magenta are *A. m. pseudomajus*, names in yellow are *A. m. striatum*.

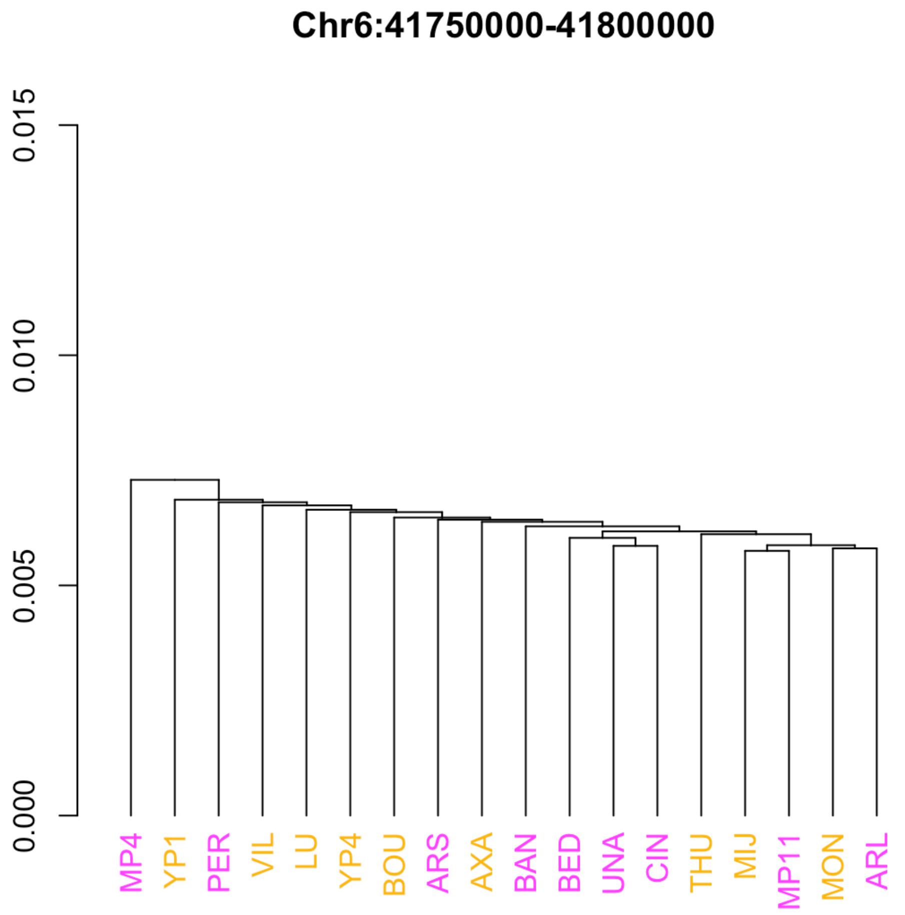
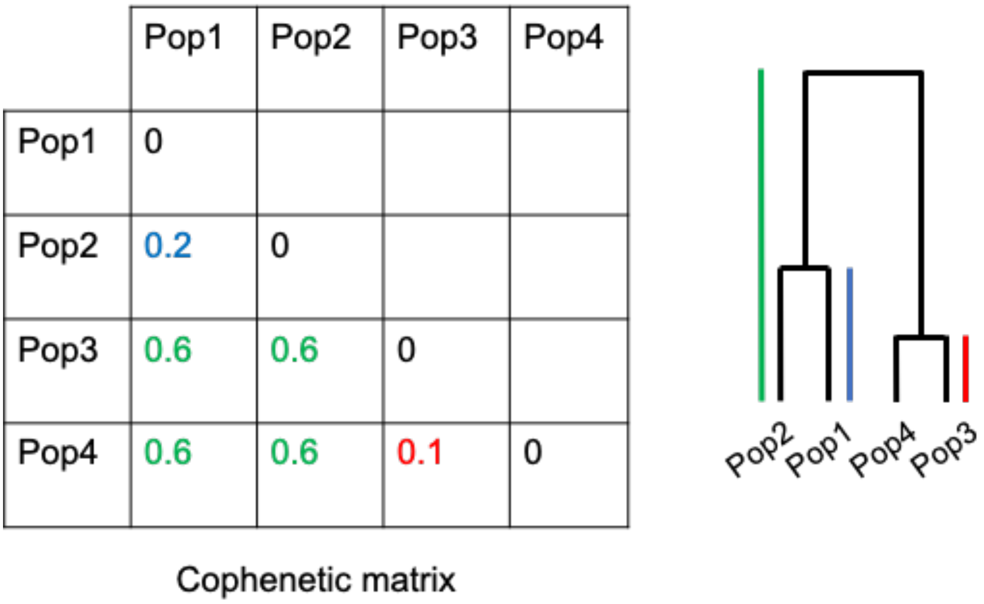


Figure 3.7: D_{XY} tree for a random genomic window

Mean D_{XY} tree for a randomly sampled 50 kb window of genomic sequence. Population names in magenta are *A. m. pseudomajus*, names in yellow are *A. m. striatum*.

A second grouping tree scan was carried out, and forest sizes were compared. About 22 % of trees were assigned to a single forest, and the four largest forests accounted for about 44 % of all genomic trees. The remaining forests were mainly small, with a mean size of 34, suggesting that many topologies show a very limited genomic distribution.



	Pop1	Pop2	Pop3	Pop4
Pop1	0			
Pop2	0.2	0		
Pop3	0.6	0.6	0	
Pop4	0.6	0.6	0.1	0

Cophenetic matrix

Figure 3.8: Populating a cophenetic matrix from an UPGMA tree

Diagram explaining how values in a cophenetic matrix are derived from an UPGMA tree. A cophenetic matrix contains minimum merging distances between all pairs of populations in a tree. The smallest value corresponds to the smallest between-population distance (red bar). The UPGMA algorithm merges subsequent populations across increasing distances (blue bar). The largest value within the matrix corresponds to the height of the tree, where the two outermost clusters merge (green bar).

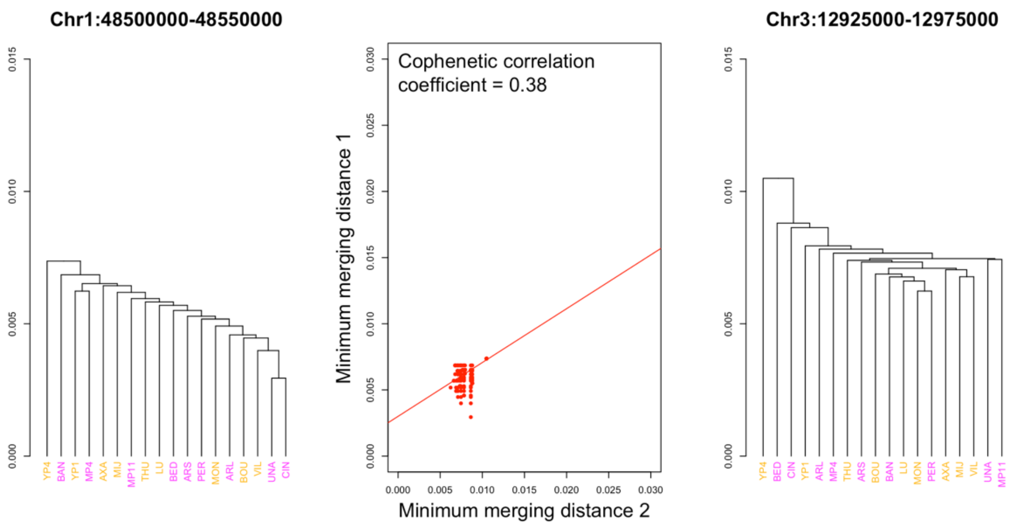


Figure 3.9: Calculation of cophenetic correlation coefficient

Calculation of the cophenetic correlation coefficient from minimum merging distances for two randomly sampled trees. The central plot shows the minimum merging distances for all pairs of populations, from each tree. An upwards-trending best fit line reflects a positive cophenetic correlation. The value of the cophenetic correlation coefficient reflects how tightly the points cluster around the fit line.

To identify D_{XY} trees with deep divisions between two groups of populations, indicative of reduced gene flow, the length of the shortest root branch (*SRB*) was calculated for each tree (Figure 3.10). A high value of *SRB* indicates a deeply rooted tree. *SRB* length was used in preference to the sum of root branch lengths to exclude groupings in which only one population was an outlier. In such cases, the longest root branch is equal to the total height of the tree (Figure 3.10). For each forest, the mean *SRB* length of all its trees was calculated. Figure 3.11 shows the relationship between forest size, and mean forest *SRB* length for one run of the tree classification algorithm. One forest was an outlier, with a mean *SRB* about 6 times greater than the mean forest *SRB*. This outlier forest contained 62 trees (0.32 % of the total number) derived from windows on chromosomes 1, 2, 4, 5, and 6.

To determine whether the outlier forest was a consistent feature of all runs, a bootstrapping approach was used. This involved running the forest classification algorithm repeatedly and recording window coordinates for all trees within the forest showing the highest mean *SRB*. Forests with fewer than four trees were ignored. Figure 3.12 shows the results for 25 bootstrap replicates. 446 windows occurred at least once in an outlier forest, but only 54 windows (0.44 % of all genomic windows), appeared in more than 50 % of replicates. These 54 windows mapped to six chromosome regions, *I-VI*, and showed an elevated mean D_{XY} compared to the rest of the genome (Figure 3.13) (*t-test*, $p < 2.2 \times 10^{-16}$). There was also evidence of reduced mean π_W in *A. m. striatum* populations (*t-test*, $p = 9.5 \times 10^{-4}$)

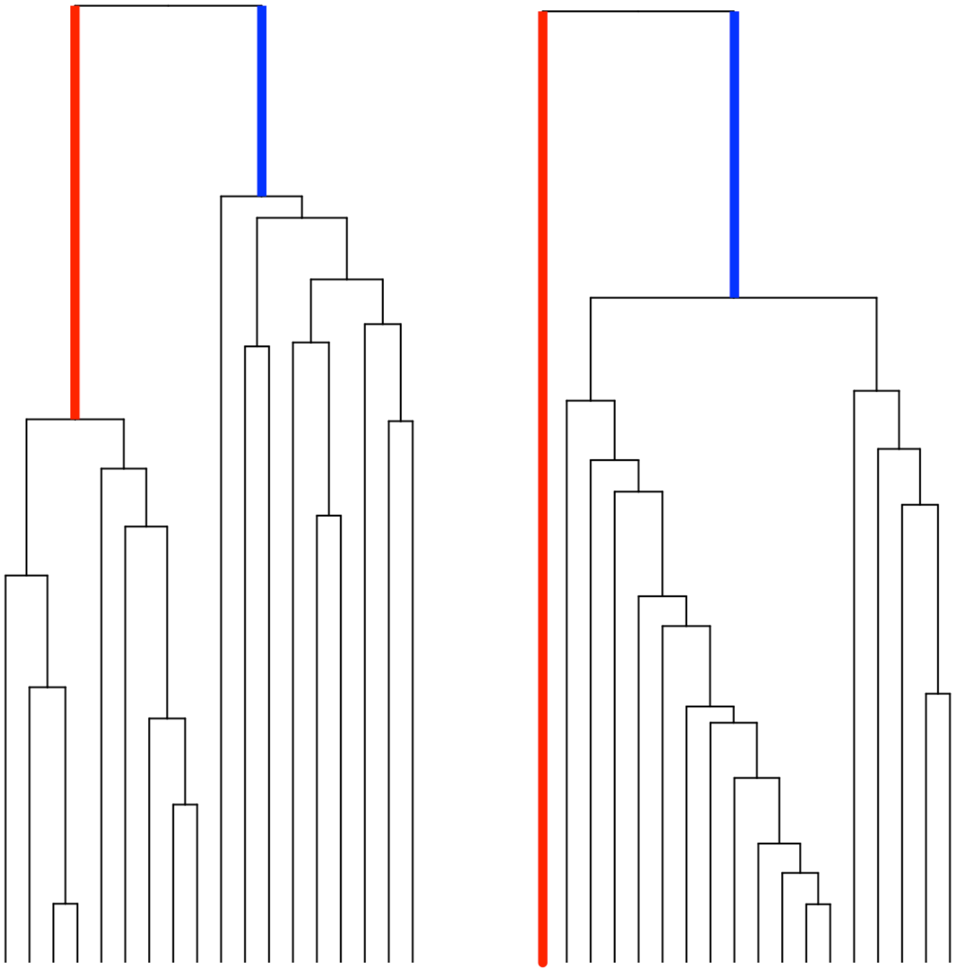


Figure 3.10: Longest- and Shortest Root Branch

Illustration of longest root branch (red) and shortest root branch (blue) on two trees. Note that, where one population is an outlier, the longest root branch is equal to the tree height.

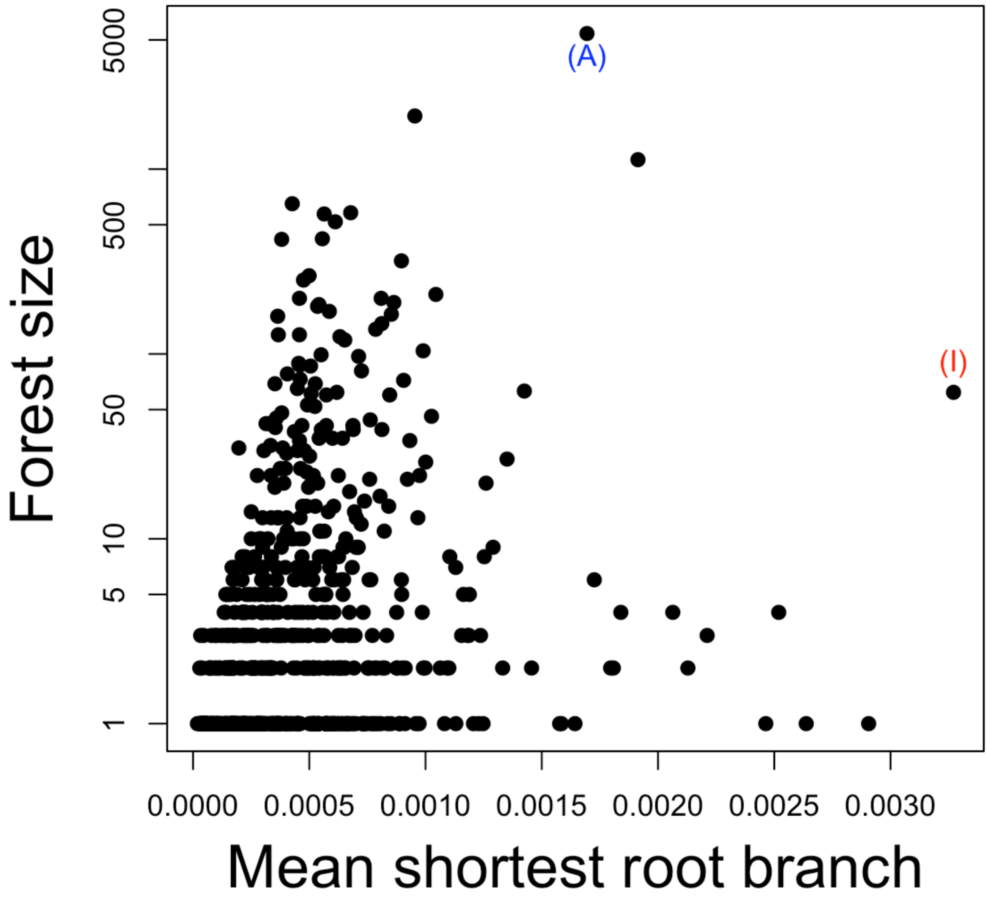


Figure 3.11: Forest size and SRB from one grouping tree scan

Summary of results from one run of the grouping tree scan method. Forest size, or number of trees within each forest, is plotted against the mean SRB of all trees in a given forest. The largest forest is denoted by (A). The forest with the greatest mean SRB is denoted by (I). The y-axis is on a logarithmic scale.

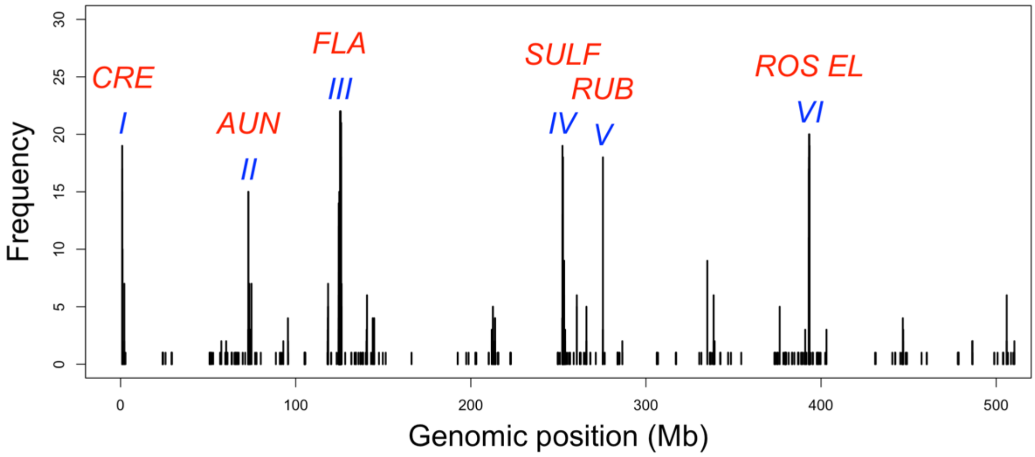


Figure 3.12: Bootstrap frequencies of genomic regions in the outlier forest

Frequency with which different genomic regions occurred within the most outlying forest (based on mean *SRB*) across 25 grouping tree scan replicates. Each labelled peak corresponds to a region that consistently appeared within bootstrap replicates (Frequency > 12). Proposed names of these regions are shown in red.

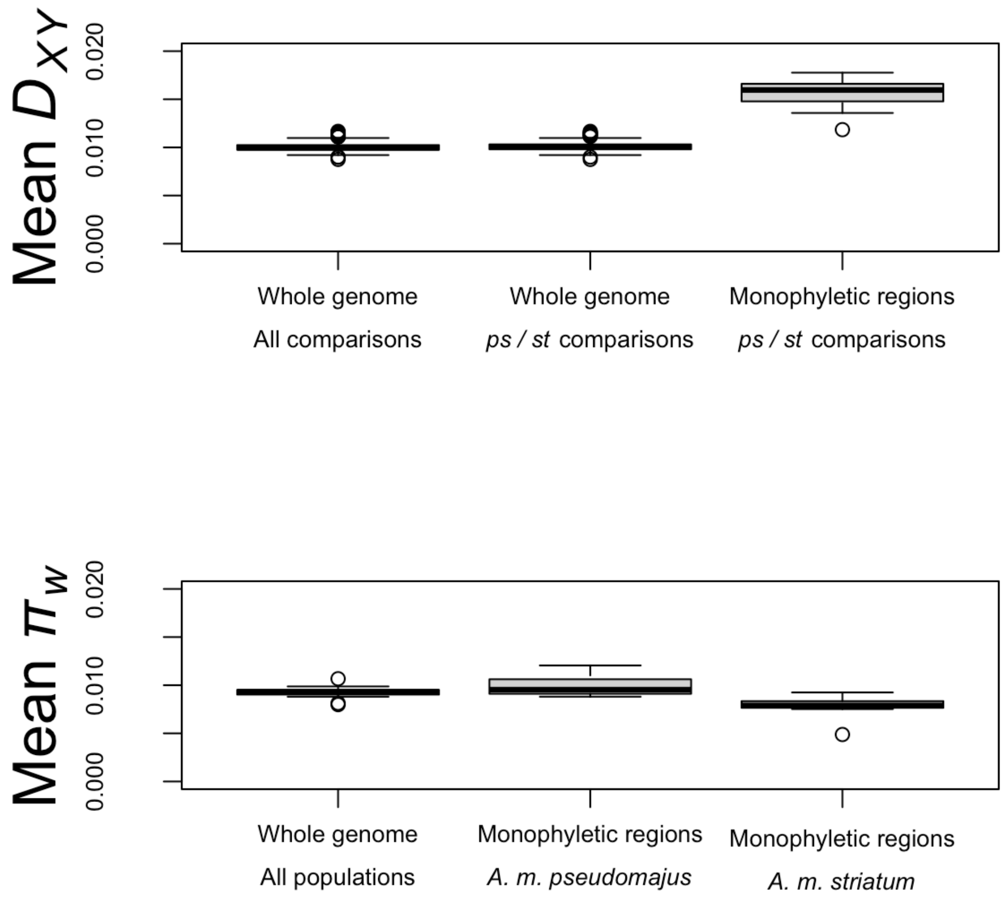


Figure 3.13: Boxplots of D_{XY} and π_w across the whole genome, and monophyletic regions

Boxplots summarising mean D_{XY} (top) and π_W (bottom) across the whole genome, compared to windows within monophyletic regions. *ps* and *st* refer to *A. m. pseudomajus* and *A. m. striatum* populations respectively. Mean D_{XY} is significantly higher within monophyletic regions than the whole genomic average calculated from all comparisons, and from only *A. m. pseudomajus* and *A. m. striatum* comparisons (t-test, $p < 2.2 \times 10^{-16}$ in both cases). Mean π_W is also significantly reduced within *A. m. striatum* populations compared to the mean of all populations (t-test, $p = 6.1 \times 10^{-3}$) and *A. m. pseudomajus* populations (t-test, p = 9.5×10^{-4}).

Results: The six regions give deeply rooted monophyletic D_{XY} trees

To determine how populations are classified by the trees in the outlier forest, mean trees for the largest forest (*A* in Figure 3.11) and outlier forest (*I* in Figure 3.11) were compared for one of the runs. The mean tree for the largest forest was shallowly rooted and gave a polyphyletic grouping (Figure 3.14, right), similar to that observed for the whole genome tree. By contrast, the mean tree of the outlier forest was deeply rooted, and gave a monophyletic grouping for *A. m. pseudomajus* and a near-monophyletic grouping for *A. m. striatum* (Figure 3.14, left). The exceptional population was *A. m. striatum* YP1, which grouped closest to the MP4 population of *A. m. pseudomajus*. YP1 and MP4 are around 4.2 km apart, and derive from opposite flanks of a natural hybrid zone. The grouping of YP1 and MP4 may thus reflect extensive gene flow between these two populations.

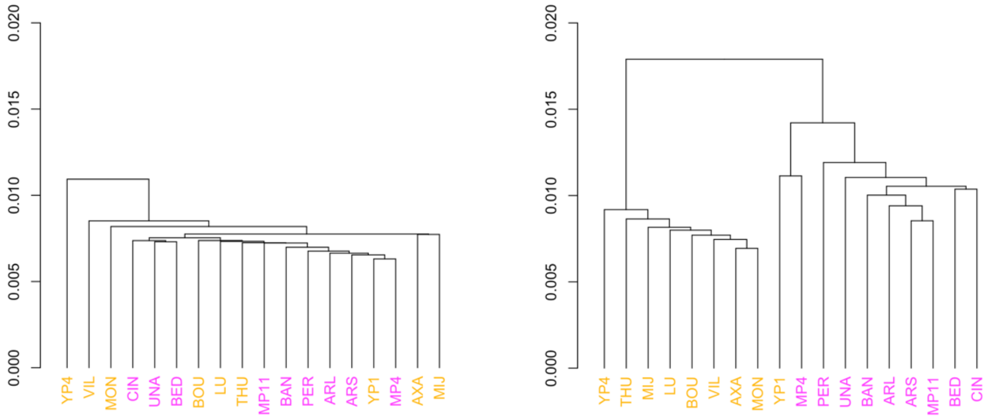


Figure 3.14: Mean D_{XY} trees of the largest and most outlying forests

Mean D_{XY} trees showing the largest forest (left) and the highest mean SRB forest (right), from the initial grouping tree scan.

To evaluate the topology of trees that consistently fall in the outlier forest, the mean D_{XY} tree for the 54 windows with this property was determined. This tree subdivided the subspecies in the same way as the mean tree for the outlier forest (Figure 3.15a). The extent to which all 54 trees gave the same subdivision was determined by classifying trees according to how their primary root divided the subspecies. This classification showed that the most abundant class (24/54) gave a monophyletic grouping for both subspecies (Table 3.1). The second most abundant (15/54) gave the same grouping as in the mean outlier forest (Figure 3.15b). The third (10/54) gave a monophyletic grouping for *A. m. striatum* but grouped MP4 of *A. m. pseudomajus* with *A. m. striatum* (Figure 3.15c). Three of the remaining five trees were monophyletic for *A. m. pseudomajus*. Thus, 68.5 % of trees that consistently belonged to the outlier forest gave monophyletic grouping for one or both subspecies.

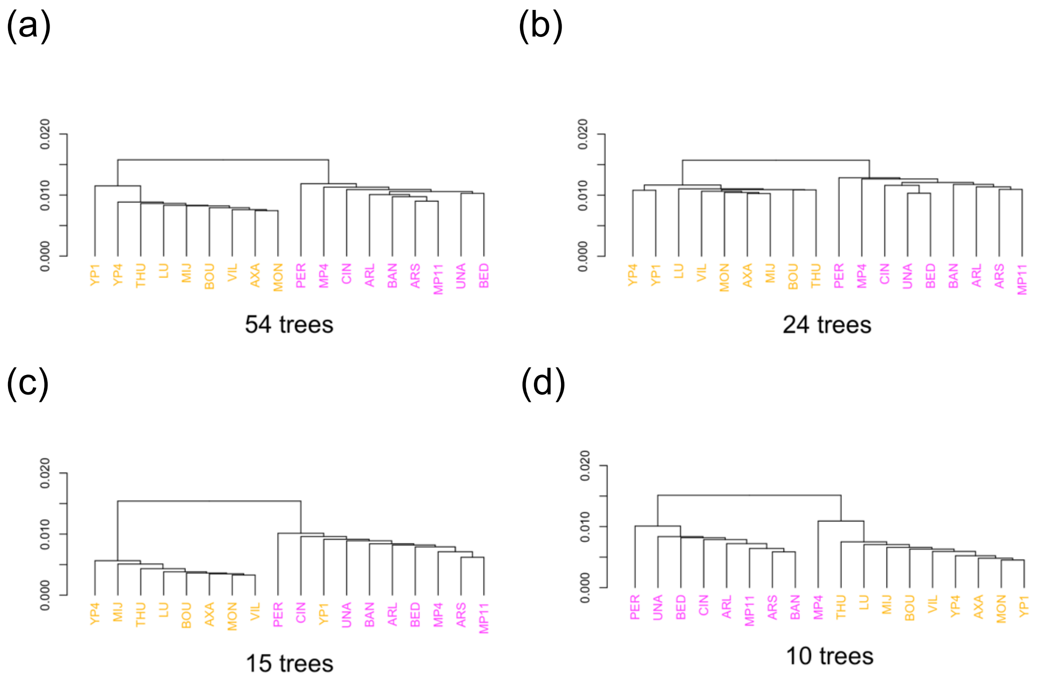


Figure 3.15: Tree topologies from the outlier forest

(a) Mean D_{XY} tree topology of all 54 windows that were consistent members of the outlier forest (> 12 bootstrap replicates). (b) Most common D_{XY} tree topology amongst the 54 windows. (c) Second most common D_{XY} tree topology amongst the 54 windows. (d) Third most common D_{XY} tree topology amongst the 54 windows. (b), (c), and (d) correspond to the first three trees in Table 3.1.

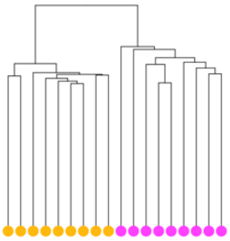
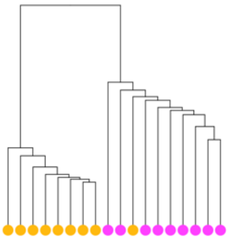
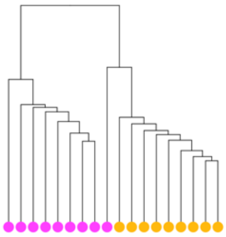
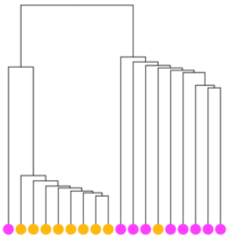
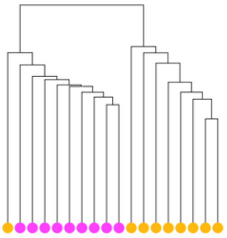
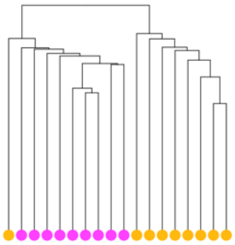
Topology	Frequency	Description
	24	Monophyletic for *A. m. pseudomajus* and *A. m. striatum*
	15	Polyphyletic. YP1 within *A. m. pseudomajus* clade
	10	Monophyletic for *A. m.* striatum. MP4 adjacent to *A. m. striatum* clade.
	2	Polyphyletic. YP1 within *A. m. pseudomajus* clade. PER adjacent to *A. m. striatum* clade.
	2	Monophyletic for *A. m. pseudomajus*. THU adjacent to *A. m. pseudomajus* clade.
	1	Monophyletic for *A. m. pseudomajus*. BOU adjacent to *A. m. pseudomajus* clade.

Table 3.1 (previous page): Consistent tree topologies within the outlier forest

Topologies, frequencies, and descriptions of the six distinct trees represented within the 54 consistent outlier windows. Magenta tips within the trees correspond to *A. m. pseudomajus* populations, and yellow tips correspond to *A. m. striatum* populations.

The six chromosomal regions giving rise to the 54 trees identified through bootstrapping will henceforth be termed monophyletic regions. Several regions contained multiple windows separated by small gaps. These additional windows were included in the analyses that follow. Therefore, monophyletic regions consist of 85 50 kb windows, comprising 2.38 Mb, or 0.47 % of the genome.

Results: Monophyletic regions harbour loci affecting flower colour variation between subspecies

The phenotypic distinction between *A. m. pseudomajus* and *A. m. striatum* is based on flower colour pattern. Two of the monophyletic regions harbour flower colour loci - region *VI* includes the *ROS* and *EL* MYB-like genes that affect magenta colour (Tavares *et al*., 2018), and region *IV* includes the sRNA locus *SULF* that affects yellow pigmentation (Bradley et al., 2017). A further monophyletic region, *II*, has recently been shown to also harbour the flower colour locus, *FLA* (Bradley *et al*., manuscript in preparation). The remaining three monophyletic regions may therefore harbour previously unidentified loci influencing flower colour, or loci that affect other traits. Two islands, *I* and *V*, were tested for phenotypic effects on flower colour. To do this, Desmond Bradley and Lucy Copsey generated F_2 hybrid populations by crossing *A. m. pseudomajus* and *A. m. striatum* individuals grown from seed collected near the hybrid zone at Planoles. Desmond Bradley then carried out genotyping work on plants. Genotypes for the monophyletic regions were determined by KASP / AFLP, using SNPs that were distinctive between *A. m. pseudomajus* and *A. m. striatum.* To minimise confounding effects of other loci known to affect flower colour, plants were also genotyped for *ROS, EL, SULF* and *FLA*. Visual inspection of flowers from plants segregating for region *V* SNPs revealed significant variation in magenta intensity, even on a *ROS* background (Figure 3.16b), suggesting that region *V* may harbour a flower colour locus. To test whether region *I* alleles show a distinct phenotypic effect, successive sib-crosses were carried out to generate an F_4 family segregating for SNPs within region *I* (pedigree shown in Appendix 2). To observe whether allelic difference at monophyletic regions resulted in distinctive colour differences, flowers were harvested from plants that were homozygous for *A. m. striatum* alleles at yellow and magenta colour genes, and photographed. By comparing flowers showing different genotypes for region *I* alleles, effects on colour could be observed. Figure 3.16a shows representative photographs of flowers showing allelic variation within regions *I* against a consistent colour gene background. Together, the results presented in Figure 3.16 show preliminary evidence that these regions may be involved in colour variation. Region *I* likely contains a locus that increases yellow in *A. m. striatum* (provisionally named the *CREMOSA* or *CRE* locus), whereas region *V* likely contains a locus that enhances magenta in *A. m. pseudomajus* (provisionally called the *RUBIA* or *RUB* locus). However, larger sample sizes are required to carry out statistical tests. A preliminary analysis of region *II* (provisionally called the *AURINA* or *AUN* locus) by Desmond Bradley has also demonstrated that the *A. m. striatum* allele results in increased yellow intensity (Bradley *et al.*, preliminary data).

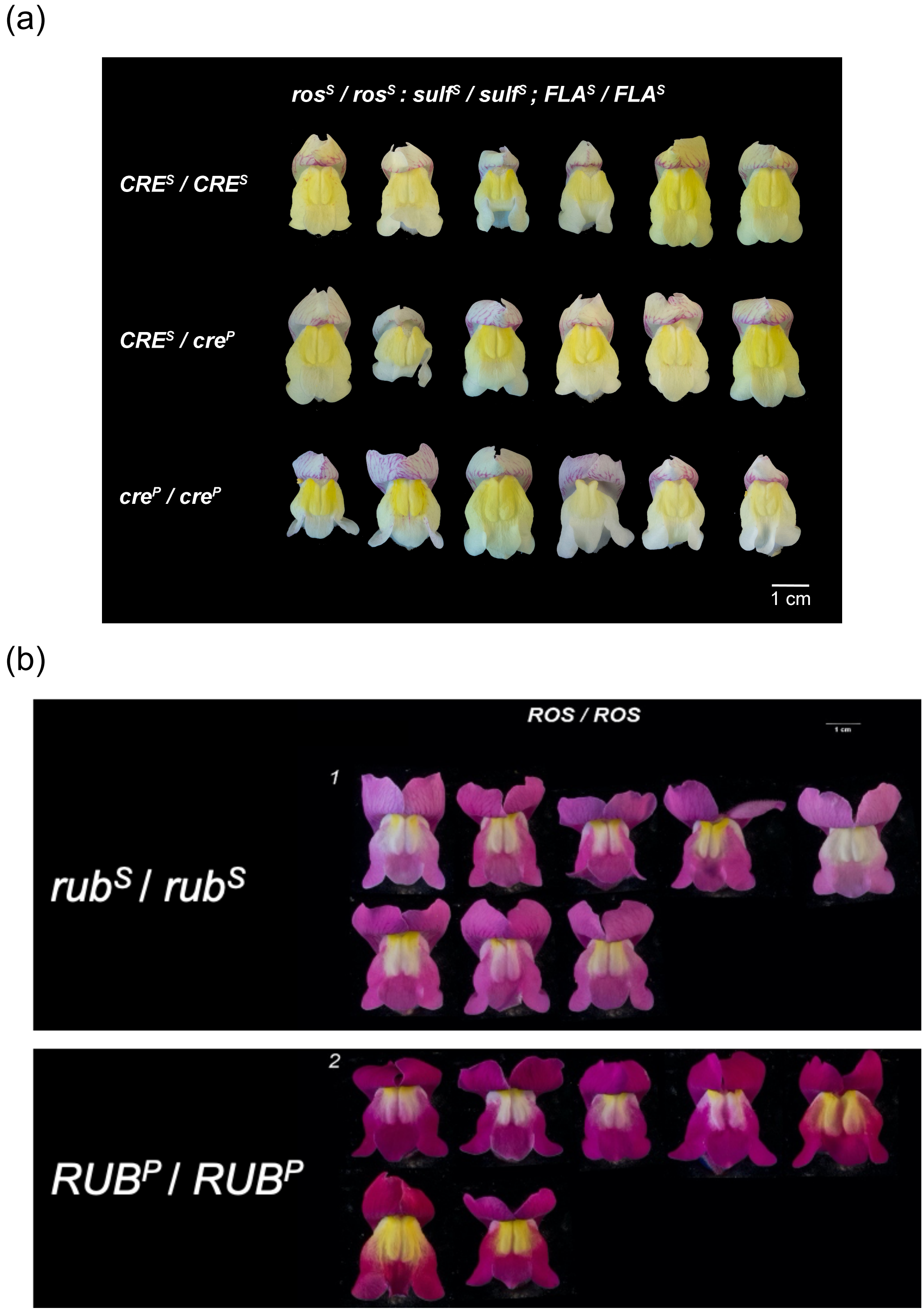

(a)
ros^S / ros^S : sulf^S / sulf^S ; FLA^S / FLA^S
CRE^S / CRE^S
CRE^S / cre^P
cre^P / cre^P
1 cm
(b)
ROS / ROS
1 cm
1
rub^S / rub^S
2
RUB^P / RUB^P

Figure 3.16 (previous page): Flower photographs showing effects of allelic variation in CRE and RUB

Phenotypic effects of *A. m. pseudomajus* and *A. m. striatum* alleles within the *CRE* and *RUB* regions. (a) Flowers from F_4 plants that are homozygous for the *A. m. striatum CRE* allele (top row), heterozygous (middle row), and homozygous for the *A. m. pseudomajus cre* allele (bottom row). The genotypes of relevant colour genes is reported at the top of the image – homozygous for *A. m. striatum ros*, *sulf*, and *FLA*. (b) Flowers from F_2 plants that are homozygous for the *A. m. striatum RUB* allele (top row), and homozygous for the *A. m. pseudomajus rub* allele (bottom row). All plants are homozygous for *A. m. pseudomajus ROS*. This work was carried out by Desmond Bradley.

Results: Identification of candidate genes underlying flower colour variation

To identify candidate genes underlying *RUB*, *AUN* and *CRE* phenotypes, differential expression analysis was carried out. mRNA was extracted from flower buds of two *A. m. pseudomajus* and three *A. m. striatum* accessions by Desmond Bradley. RNAs were sequenced by Yongbiao Xue. Annabel Whibley aligned RNAseq data to the reference genome, and carried out differential expression analysis using DESeq2 (Love, Huber, and Anders, 2014). I then analysed output of the differential expression analysis to identify differentially expressed genes within monophyletic regions. Differential expression *p* values obtained from DESeq2 are corrected for multiple testing, and will therefore be referred to as *p*-adjusted values. Of the 45,648 genes predicted in the reference genome, 1,725 (3.8 %) were differentially expressed between *A. m. pseudomajus* and *A. m. striatum* (*p*-adjusted < 0.01). The monophyletic regions (85 windows) contained 968 predicted coding sequences (CDSs), 29 (3 %) of which showed differential expression. These 29 genes were distributed across all six monophyletic regions. Four of the 29 genes corresponded to *ROS1*, *ROS2*, *ELUTA* (region *VI*) and *FLA* (region *III*). Six differentially expressed genes were within region *IV*, but these did not include *SULF*, a small RNA (sRNA) producing inverted repeat locus involved in repressing yellow pigmentation. In the absence of sRNA data, it is unclear whether this is due to non-differential *SULF* expression or insensitivity of mRNAseq to sRNA expression. There were however two genes showing high similarity to *Am4'CGT,* suggesting that region *IV* may be involved in yellow pigment regulation through elements other than *SULF*, or that *SULF* acts on these loci as well as its previously defined target on chromosome two. Region *II* included three differential genes, one of which corresponded to *AmAS1*, which encodes aureusidin synthase, the enzyme responsible for synthesising yellow aurone pigments from chalcones (Nakayama *et al*., 2000). Thus, *AmAS1* is a strong candidate for the yellow flower locus, *AUN.* The only differentially expressed gene within region *V* encoded a flavonol synthase, which has been implicated in controlling the concentration of magenta anthocyanin pigment through competitive substrate utilisation (Luo *et al.*, 2016). This gene is therefore a strong candidate for magenta flower locus, *RUB*. Region *I* contained two differential genes, encoding an *O*-methyltransferase, and a pyrophosphorylase (based on NCBI BLASTX similarity). The activity of *O*-methyltransferases has been implicated in flower colour, through interactions with the anthocyanin biosynthesis pathway (Akita *et al.*, 2011, Du *et al.,* 2015, Okitsu *et al.*, 2018). Thus, it is possible that this gene corresponds to the yellow flower locus, *CRE*, and might affect yellow through direct methylation of aurones, or by affecting flux of substrates that are shared between the anthocyanin and yellow pathways. These analyses identify candidate loci for *RUB*, *AUN* and *CRE* but further expression and genetic tests would be needed to confirm their assignment. It is also possible that one or more of the monophyletic regions include loci that influence flower colour through changes that do not modify mRNA transcript levels, such as sRNA loci.

Results: Relationship between monophyletic regions and other measures of population variation

The above genome scans of D_{XY} trees have identified six regions with elevated D_{XY} that likely harbour barriers to gene flow between subspecies. To see whether these regions show visible elevation in pairwise genome scans, D_{XY} and F_{ST} values were averaged across all comparisons, and plotted (Figure 3.17). No specific loci stand out within the D_{XY} scans compared to the background D_{XY} (Figure 3.17 (top)). However, three of the six monophyletic regions do show strong peaks in F_{ST} scans (Figure 3.17 (bottom)).

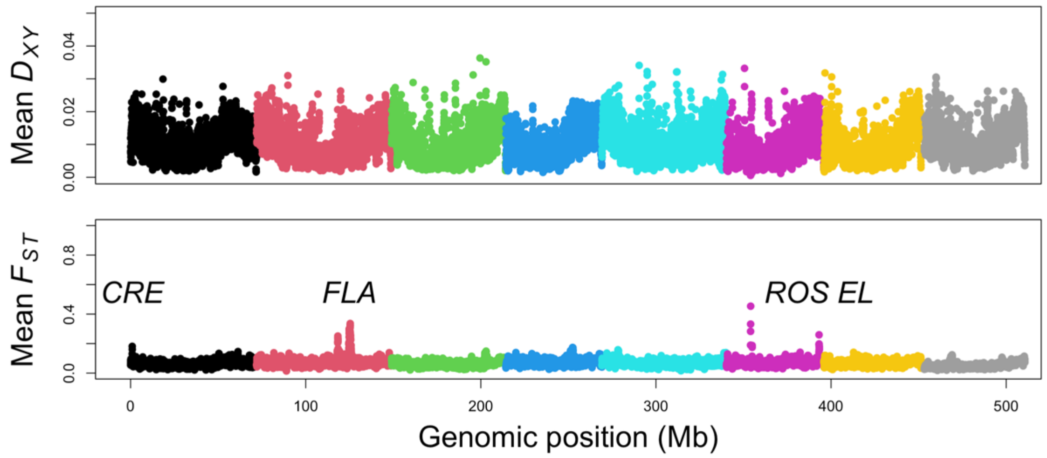


Figure 3.17: Mean D_{XY} and F_{ST} from all A. m. pseudomajus / A. m. striatum comparisons

Mean D_{XY} (top) and F_{ST} (bottom) from all population comparisons, across all 19,520 overlapping 50 kb genomic windows. The *CRE*, *FLA* and *ROS EL* regions show local F_{ST} peaks. Sequential changes in colour reflect different chromosomes (1-8). Two other distinct peaks, on chromosome 2 and chromosome 6, are uncharacterised (the leftmost F_{ST} peak on chromosome 6 does not correspond to *ROS* or *EL*).

To clarify the reasons for these observations, 1 kb window scans were carried out centred on the six monophyletic regions. Figures 3.18 - 3.20 show plots of D_{XY}, π_W, D, and F_{ST} across monophyletic regions, and 1 Mb flanking regions. For F_{ST} and D_{XY}, average values were plotted, based on all pairwise comparisons between *A. m. pseudomajus* and *A. m. striatum*. π_W was averaged across all populations. D was calculated by subtracting mean population π_W from the mean D_{XY} of all comparisons. To estimate the proportion of SNPs within the monophyletic region (and surrounding regions) that show monophyletic distributions, trees were constructed for all 1 kb windows. The locations of trees showing either of the two most frequent monophyletic topologies in Table 3.1 were marked (Figures 3.18 - 3.20, red and purple asterisks). The second most common topology was also marked (Figures 3.18 - 3.20 , blue asterisks). Across all islands, signatures of contiguous monophyletic trees were detected, illustrated by adjacent asterisks. I will refer to these contiguous regions as monophyletic islands.

Within all regions except *FLA*, the most commonly detected topology from the outlier forest was the most common, doubly monophyletic, topology. The *CRE* and *AUN* monophyletic islands were both small (< 50 kb). Both islands coincided with a small F_{ST} and D peak, suggesting that relative divergence was due to increased D_{XY} rather than reduced π_W. The largest monophyletic island, at the *FLA* region (Figure 3.19a), appeared to span over 500 kb. A larger polyphyletic island, comprising trees which group the *A. m. striatum* YP1 population within the *A. m. pseudomajus* clade, was detected immediately adjacent to the monophyletic island. Most of the regions showed elevated D_{XY} compared to π_W. The *SULF* region (Figure 3.19b) showed much missing data, reflecting low read depth in one or more populations. A small monophyletic island was detected to the left of the low coverage region. This demonstrated that the *SULF* region is only detectable as a 50 kb monophyletic region due to monophyletic SNPs within adjacent windows. This likely reflects the absence of *SULF* within populations of *A. m. striatum* (Bradley *et al.*, 2017). The *RUB* monophyletic island spanned the full width of the 50 kb monophyletic region. No F_{ST} peak was visible, although both D_{XY} and D were slightly elevated. The *ROS EL* monophyletic island was the second largest, spanning around 250 kb, although the island showed small gaps between F_{ST} peaks. The left-hand F_{ST} peak at the *ROS EL* monophyletic region showed no corresponding D peak. This demonstrates that locally elevated F_{ST} is due to reduced π_W, rather than increased D_{XY}. The left-hand F_{ST} peak corresponds to the location of the *ROS1* and *ROS2* genes (Tavares *et al.*, 2018).

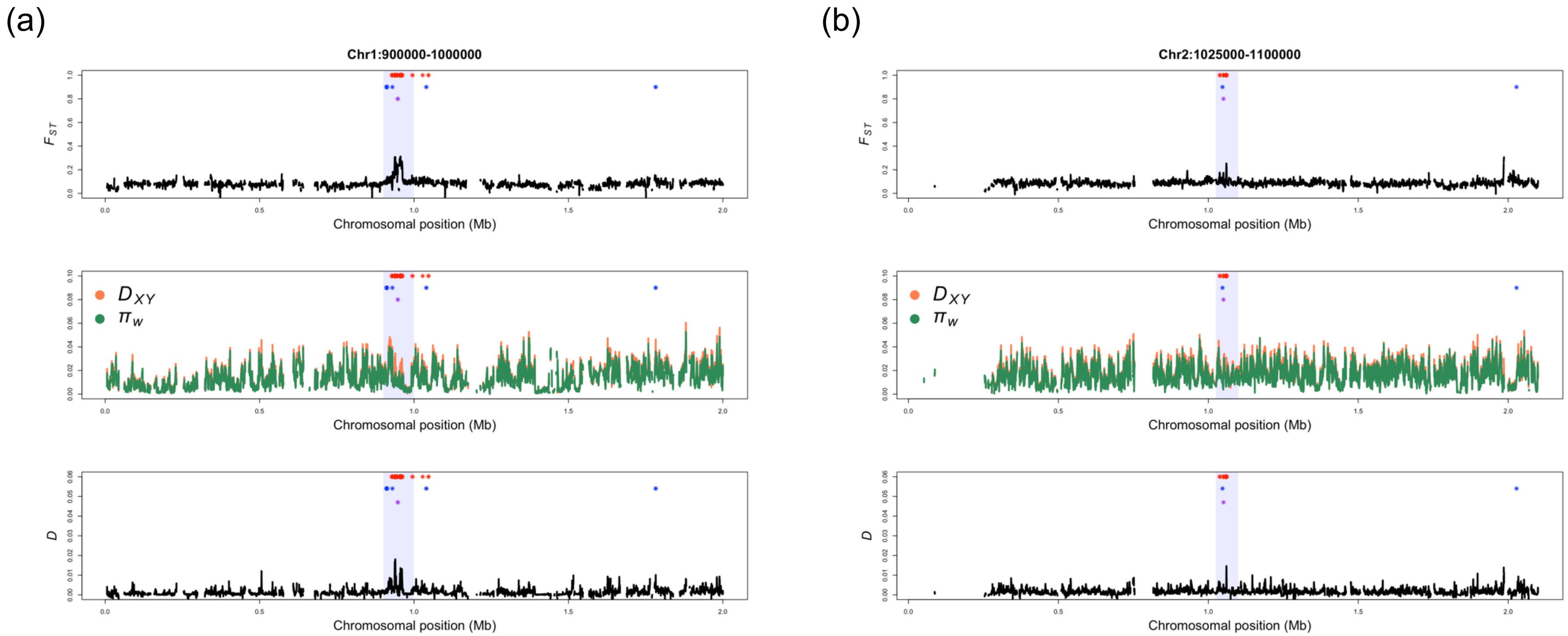


Figure 3.18: 1 kb window plots across the CRE and AUN monophyletic regions

F_{ST}, D_{XY}, and D, plotted across (a) the 100 kb *CRE* monophyletic region with 1 Mb flanks, and (b) the 75 kb *AUN* monophyletic region, with 1 Mb flanks. F_{ST} and D_{XY} are averaged from all population comparisons, and π_w is averaged from all populations. The blue region represents the monophyletic region identified in the grouping tree scan. Asterisks above the plot denote 1 kb windows that yield any of the top three topologies from Table 3.1. Red asterisks show the most common (doubly monophyletic) topology. Blue asterisks show the second most common (polyphyletic) topology, and purple asterisks show the third most common (*A. m. striatum* monophyletic) topology.

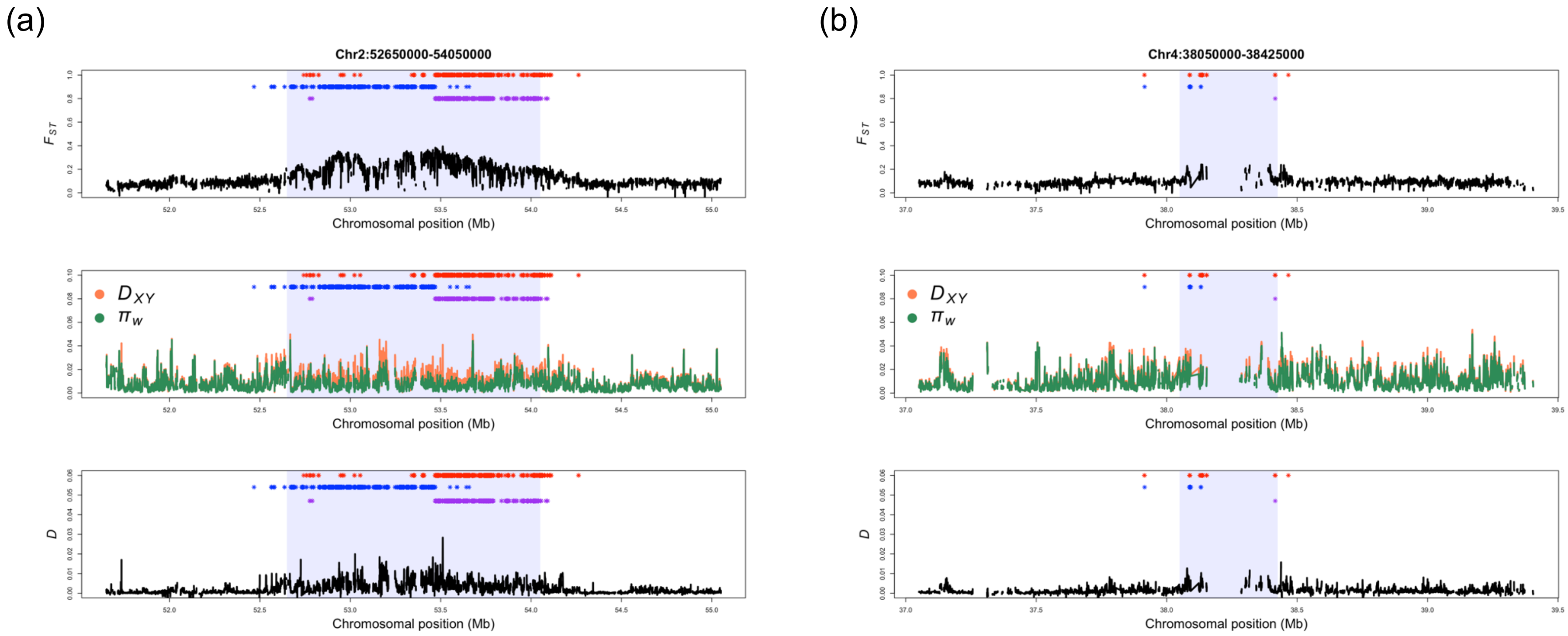


Figure 3.19: 1 kb window plots across the FLA and SULF monophyletic regions

F_{ST}, D_{XY}, and D, plotted across (a) the 1.4 Mb *FLA* monophyletic region with 1 Mb flanks, and (b) the 375 kb *SULF* monophyletic region, with 1 Mb flanks. F_{ST} and D_{XY} are averaged from all population comparisons, and π_W is averaged from all populations. The blue region represents the monophyletic region identified in the grouping tree scan. Asterisks above the plot denote 1 kb windows that yield any of the top three topologies from Table 3.1. Red asterisks show the most common (doubly monophyletic) topology. Blue asterisks show the second most common (polyphyletic) topology, and purple asterisks show the third most common (*A. m. striatum* monophyletic) topology.

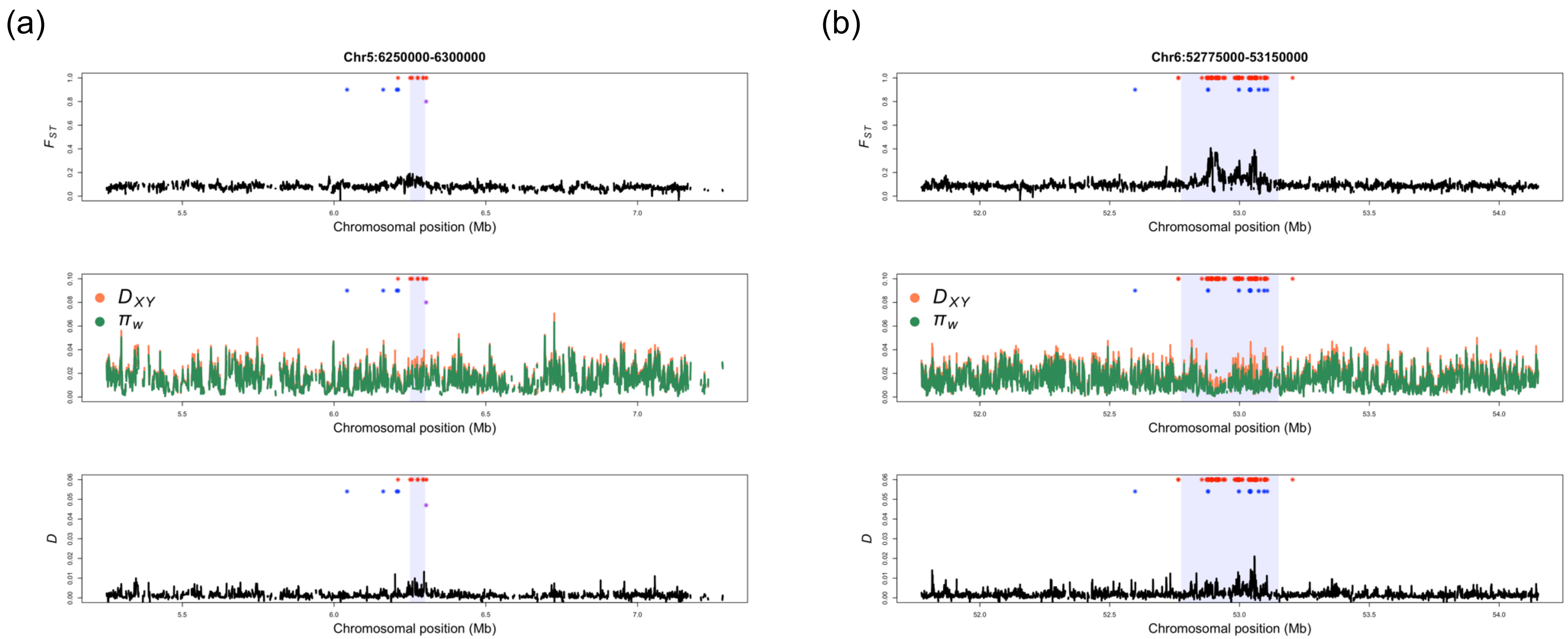


Figure 3.20: 1 kb window plots across the RUB and ROS EL monophyletic regions

F_{ST}, D_{XY}, and D, plotted across (a) the 50 kb *RUB* monophyletic region with 1 Mb flanks, and (b) the 375 kb *ROS EL* monophyletic region, with 1 Mb flanks. F_{ST} and D_{XY} are averaged from all population comparisons, and π_W is averaged from all populations. The blue region represents the monophyletic region identified in the grouping tree scan. Asterisks above the plot denote 1 kb windows that yield any of the top three topologies from Table 3.1. Red asterisks show the most common (doubly monophyletic) topology. Blue asterisks show the second most common (polyphyletic) topology, and purple asterisks show the third most common (*A. m. striatum* monophyletic) topology.

Discussion

Grouping tree scans reveal hidden islands of genomic divergence

The aim of this work was to use tree classification of genomic windows scans to identify consistent barriers to gene flow between *A. m. pseudomajus* and *A. m. striatum*, independently of phenotypic observations. Applying phylogenetic approaches to study D_{XY} patterns addresses some of the challenges of interpreting genome-wide datasets. Multiple population comparisons, compared to two-way comparisons, have greater power to differentiate between consistent and sporadic D_{XY} elevation (Seehausen et al., 2014). This increases sensitivity to consistent signals that may not be individually strong. It also reduces the effect of noise arising from ancestral π_W, which reflects a range of historical and demographic factors and is not expected to be consistent between populations. Fundamentally, this approach makes distinct genomic regions more directly comparable. This is evidenced by the diverse nature of the candidate regions identified. For example, *FLA* resides in a region of low recombination, whereas *CRE* and *AUN* likely do not (Li et al., 2019). *ROS-EL* shows locally reduced π_W, whereas *AUN* does not. The latter observation that not all islands have reduced π_W may explain why some islands have not been detected through F_{ST} analysis.

Tavares et al. (2018) suggested that reduced π_W at *ROS-EL* was consistent with a selective-sweep-mediated origin. This reduced π_W is apparent when plotting π_W and D_{XY} across the region, being especially apparent around the first F_{ST} peak, which corresponds to the location of the *ROS1* and *ROS2* genes. A strong F_{ST} peak, along with the absence of a D peak suggest that π_W likely accounts for most of the relative divergence at *ROS1* and *ROS2*. Interestingly however, the downstream peak corresponding to the *EL* locus shows elevation in both D_{XY} and D, which is not reflective of a selective-sweep-mediated origin

Monophyletic regions contain intrinsic epistatic barrier genes

The 18 populations presented here were sampled across considerable geographical distance, and from different habitats. For the observed barriers to reflect cryptic environmental differences, such differences would have to be shared between these disparately sampled populations in a subspecies-specific manner. Divergence through environmental adaptation is therefore unlikely. The available evidence most strongly supports an intrinsic epistatic genetic barrier. Almost all of the genetic divergence between *A. m. pseudomajus* and *A. m. striatum* is accounted for by six genomic regions. These six regions show concordant monophyletic tree topologies, and each contains at least one candidate gene pertaining to a single trait: flower colour. If the genetic barrier had an additive effect on fitness through heterozygote disadvantage then adaptation would likely make use of traits other than colour. In this case, elevated SRB should be detectable in a greater diversity of trees.

More evidence for an intrinsic epistatic barrier has been provided through cline analyses. To date, four of the six monophyletic regions detected here have been shown to have associated steep clines between *A. m. pseudomajus* and *A. m. striatum* across the Planoles hybrid zone (Bradley *et al.*, 2017, Tavares *et al.*, 2018, Bradley *et al.*, preliminary data, Field *et al.*, preliminary data). Steep clines, contrasting against the shallow clines seen across most of the genome, reflect genomic regions where allele frequency boundaries between populations are sharply defined. This provides strong, direct evidence of barriers to gene flow. Coincident clines between at least 4/6 monophyletic regions suggests that shared evolutionary forces are acting on them, providing evidence that genes may be coadapted (Barton and Hewitt, 1985).

The origin of intrinsic epistatic barriers

The detection of signatures of consistently elevated D_{XY} at only a small subset of genomic loci suggests that *A. m. pseudomajus* and *A. m. striatum* underwent a period of historical isolation, before coming into secondary contact. Whilst in isolation, divergence through mutation and drift caused D_{XY} to increase across the whole genome. Alleles conferring adaptive advantages were able to sweep to fixation within populations. Upon secondary contact, the two subspecies underwent extensive gene flow, but this was resisted at barrier gene regions with intrinsic epistatic interactions. Areas of reduced π_w within barrier regions may represent signatures of historical selective sweeps. Alternatively, elevated D_{XY} in monophyletic regions may not reflect resistance to gene flow, but instead reflect recruitment of ancient alleles affecting flower colour, perhaps originating through introgression from other species. According to this view, the ancient alleles would have undergone recombination while they were polymorphic in the population. For example, if an ancient allele takes 1,000 generations to become fixed, the linked region of elevated D_{XY} would be of the order of 0.1 cM. Assuming that the rate of recombination in *Antirrhinum* is between 0.3 - 3 cM/Mbp (Tavares *et al.*, 2018), monophyletic islands would be expected to span between 30 kb and 300 kb. This is broadly consistent with observed monophyletic island sizes, except at the 500 kb *FLA* island. It is unclear, however, why ancient hybridisation would only recruit flower colour alleles.

Evolutionary relevance of colour genes

Extensive research in *Antirrhinum* (Shang et al., 2010, Bradley et al., 2017) and other plants (Lunau, Wacht, and Chittka, 1996, Schemske and Bradshaw, 1999, Reverté *et al.*, 2016) has demonstrated that pollinators show preferences for specific flower colour patterns. Pollinator attraction is essential for reproductive success in sexually reproducing plants. Therefore, genes involved in the refinement and regulation of flower colour patterns are likely subject to sexual selection (Moore and Pannell, 2011). Because sexual selection can drive very rapid evolution of genetic barriers, it is postulated that sexual traits may play a role in the early stages of speciation (Panhuis *et al.*, 2001, Aagaard et al., 2013). Such early-stage speciation around a sexual trait may be taking place within *A. m. pseudomajus* and *A. m. striatum*. Fitness may also reflect frequency-dependent selection, with pollinators having a preference for predominant local colour phenotype (Smithson and Macnair, 1996).

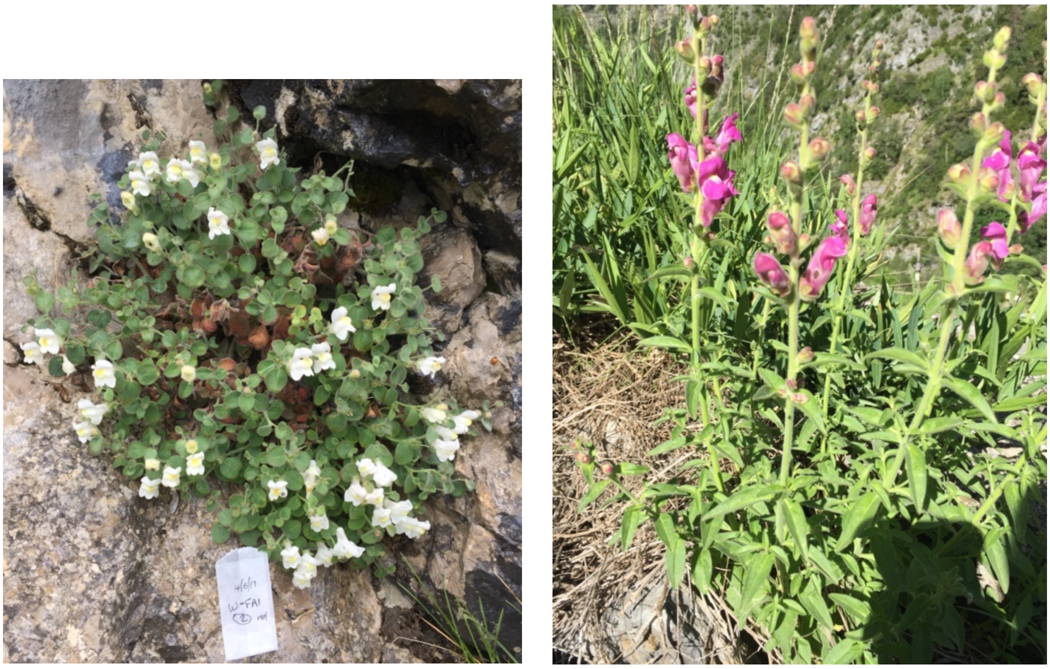
4: Monophyletic signatures reveal genetic barriers underpinning growth habit divergence in *Antirrhinum*

Introduction

Analysis of genomic divergence in *A. m. pseudomajus* and *A. m. striatum* demonstrated that barriers to gene flow co-localise with loci controlling flower colour. I hypothesise that *A. m. pseudomajus* and *A. m. striatum* are subject to intrinsic postzygotic isolation arising due to negative epistasis, with non-parental allele combinations imparting reduced fitness upon hybrid progeny. However, the perseverance of barriers to gene flow is not necessarily completely dependent on epistatic relationships. Gene flow barriers may arise around loci involved in differential adaptation to ecological niches. In such cases, gene flow will be restricted between populations that occupy different niches. Because *Antirrhinum* species show a great deal of diversity in traits other than colour, barrier gene analyses can be expanded to explore differential adaptation.

Growth habit in Antirrhinum

Perhaps the most significant phenotypic differences within the *Antirrhinum* species group are in growth habit. Growth habit differences in *Antirrhinum* encompass a range of traits, including plant height, leaf size, leaf shape, flower size, branching angles and hairiness. Two distinct habits are recognised, each reflecting a characteristic combination of these traits (Figure 4.1). In this thesis, I will refer to these habits as alpine and ruderal. The alpine habit is characterised by compact, bushy species that grow on rock faces with branches that trail along the growing surface. Plants with the ruderal habit are much more widespread, growing on sloping hills and grasslands, as well as human-disturbed sites such as roadsides. Ruderal species are tall and erect, with large leaves and flowers. The genus *Antirrhinum* has traditionally been grouped into morphological subsections. The alpine habit is reflected by subsection Kickxiella, and the ruderal habit by both Antirrhinum and Streptosepalum (Rothmaler, 1956, Webb, 1971, Sutton, 1988). Studies of allometry (correlated variation in size and shape) have demonstrated that growth habit traits are likely to be under the control of many underlying genes of individually small effect (Feng *et al.*, 2009). However, not all growth habit traits show a complex genetic architecture. Tan *et al.* (2020) recently demonstrated that hairiness, a phenotype associated with alpine species, is controlled by a single glutaredoxin gene. The extrusion of trichome hairs is implicated in protection from a number of biotic and abiotic factors. *Hairy* prevents the emergence of trichomes above the fourth leaf internode. Alpine species are typically hairy, which implies that they are homozygous for the recessive *hairy* allele. Ruderal species are generally hairless above internode four, implying heterozygosity or homozygosity for *Hairy*. Characterising the genetic basis of divergence between species showing distinct growth habits may shed light on the contribution of major effect loci, and polygenes.



Alpine

Ruderal

Figure 4.1: Examples of alpine and ruderal growth habits

Photographs of wild plants in Spain showing characteristic alpine (left) and ruderal (right) growth habits. The plant on the left is *Antirrhinum molle*, and the plant on the right is *Antirrhinum majus pseudomajus*.

Identifying growth habit barrier genes through root division analysis

In Chapter 3 I identified barrier gene regions between *A. m. pseudomajus* and *A. m. striatum* using an approach that clustered trees based on their topological similarity. This revealed that the majority of genomic divergence was accounted for by subgenomic regions giving monophyletic trees. A simpler approach to characterising divergence between the two subspecies might therefore have been to identify all subgenomic regions giving monophyletic trees. However, this would have violated the stipulation that regions should be identified independently of phenotype, and introduced ascertainment bias. Controlling for ascertainment bias when analysing *A. m. pseudomajus* and *A. m. striatum* was important in order to test the hypothesis that the reproductive barrier was underpinned by colour genes only. In comparing species with different growth habits, no such prior assumptions are held. Alpine and ruderal species are characterised as diverging in a wide variety of traits, but there is not a fixed definition between the two phenotypes. Furthermore, there is no prior hypothesis that the regions underlying growth habit should necessarily be the most divergent. Alpine and ruderal species show high genomic divergence, and a broad geographic distribution hints that many populations are unlikely to have undergone gene flow (Duran-Castillo *et al.*, 2022). In Chapter 3, dividing trees based on their primary root division was an effective means of classifying trees within the outlier forest (Table 3.1). In this chapter, I will therefore develop root division analysis as an alternative grouping tree scan. This will facilitate the direct detection of subgenomic regions showing trees that are monophyletic for alpine and ruderal growth habit, which may contain barrier genes. In practice, this resembles a much-simplified version of the *Twisst* approach (Martin and Van Belleghem, 2017), which searches for a pre-defined monophyletic tree topology rather than characterising all topologies within the genome.

Multiple evolutionary scenarios can give rise to monophyletic signatures

Patterns of diversity at barrier regions can reveal the evolutionary relationships between species. In principle, genomic regions harbouring barriers to gene flow between alpine and ruderal species will give trees that are monophyletic for both growth habits. However, barrier genes are not the only evolutionary mechanism which could give rise to monophyletic groupings. Alternatively, the shared adaptations of alpine and ruderal species may have arisen through phenotypic convergence, where evolution causes populations to become more phenotypically similar (Arendt and Reznik, 2008). A convergence hypothesis is supported by evidence from whole genome phylogenetics, which suggests that alpine species of *Antirrhinum* form multiple distinct clades within the species tree, and therefore arose separately (Wilson and Hudson, 2011, Tan *et al.*, 2020, Duran-Castillo et al., 2022). Stern (2013) proposed three evolutionary hypotheses that can underlie convergent evolution. The first, parallel evolution, postulates that distinct habits may have arisen through recruitment of pre-existing growth habit alleles that have arisen as polymorphisms within populations (standing genetic variation). Alternatively, distinct traits contributing to growth habit divergence may have arisen within different populations, and been shared through hybridisation following secondary contact between populations showing the same growth habit. Finally, convergent evolution may take place through growth habit alleles that have been inherited from a shared ancestral population. Stern proposed the collective term collateral evolution for the latter two of these hypotheses, because they both deal with variation which is inherited "from the same stock but by a different line". Here, for consistency with other chapters, I will separately refer to the third hypothesis (convergent evolution through ancestral shared variation) in terms of historical barrier genes. For the second hypothesis (sharing of alleles growth habit through hybridisation), I will use the term allele sharing.

Recruitment from standing variation, allele sharing, and historical barrier genes can all result in monophyletic patterns of allelic variation between alpine and ruderal species. However, it may be possible to differentiate between these hypotheses by measuring the size of monophyletic islands. The parallel-allele-recruitment hypothesis predicts small islands, because alleles are required to have persisted against ongoing recombination for up to five million years, the predicted age of the *Antirrhinum* species group (Vargas *et al.*, 2009). Repeated rounds of recombination are expected to break down haplotypes. After one million years (~ 1 million generations), monophyletic islands are expected to span a genetic distance in the order of 10^{-4} cM. Assuming a recombination rate of 0.3-3 cM/Mb (Tavares *et al.*, 2018), the physical distance covered by monophyletic islands should be between 30 bp and 300 bp, depending on whether they are localised to high or low recombining regions. Under an allele sharing hypothesis, growth habit alleles will initially exist as polymorphisms following hybridisation. However, because they confer a fitness advantage, they may become fixed relatively quickly through selective sweeps. The *ROS-EL* flower colour locus shows reduced local π_W, consistent with an historic selective sweep (Tavares *et al*., 2018). Reduced π_W at *ROS-EL* spans between 10-50 kb, in a region with a high rate of recombination. This value is expected to be up to ten times higher in a low recombining region. This provides an estimated monophyletic island size of 10-500 kb under the allele sharing hypothesis. If monophyletic islands arise due to barriers to gene flow, then the size of the island will depend upon how much gene flow has taken place between species, which is unclear.

Aim of this work

This work aims to identify barrier genes involved in growth habit divergence between alpine and ruderal growth habits. F_{ST} scans will first be carried out to compare relative genomic divergence landscapes between species. The presence of peaks may indicate barrier gene candidates. An absence of distinct peaks will suggest that growth habit divergence is underpinned by many loci of small effect. To directly investigate gene flow, a root division analysis will be carried out. By identifying subgenomic regions yielding D_{XY} trees that are monophyletic for both growth habits, candidate monophyletic islands of divergence can be identified. These monophyletic islands will be examined at a finer resolution to look for enrichment of SNPs giving monophyletic groupings. The size of high confidence monophyletic islands will be measured, in order to test hypotheses as to whether they arose through ancestral divergence, parallel allele recruitment, or allele sharing.

Results

Results: Mapping of pool-seq data from eight *Antirrhinum* species

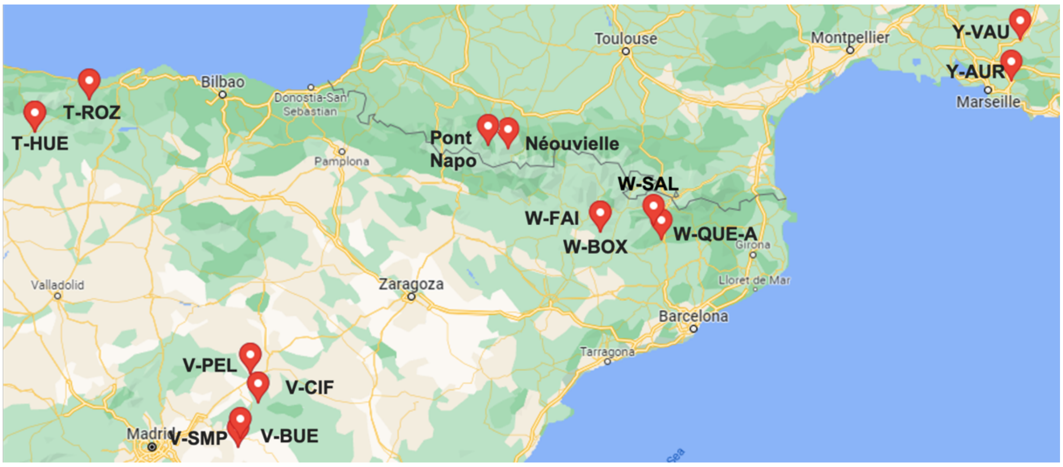


Figure 4.2: Locations of 16 sampled populations within France and Spain

Sampling locations of 16 alpine and ruderal populations. Pool details and GPS coordinates can be found in Appendix 1 and Appendix 4.

To explore the relationships between alpine and ruderal species of *Antirrhinum*, leaves were collected from 16 populations for sequencing. Populations comprised eight species, each being sampled from two distinct geographical areas (Figure 4.2) (pool details recorded in Appendix 1 and Appendix 4). *A. m. pseudomajus*, *A. m. striatum*, *A. latifolium*, and *A. braun-blanquetti* showed ruderal growth habits. *A. molle*, *A. sempervirens*, *A. microphylum*, and *A. pulverulentum* had alpine habits. Leaves were pooled and sequenced, and pool-seq data was mapped to the *Antirrhinum majus* reference genome using BWA-MEM. Appendix 4 shows genomic mapping statistics. Ruderal species showed higher genomic coverage on average compared to alpine species, reflecting divergence between alpine species and the ruderal reference genome.

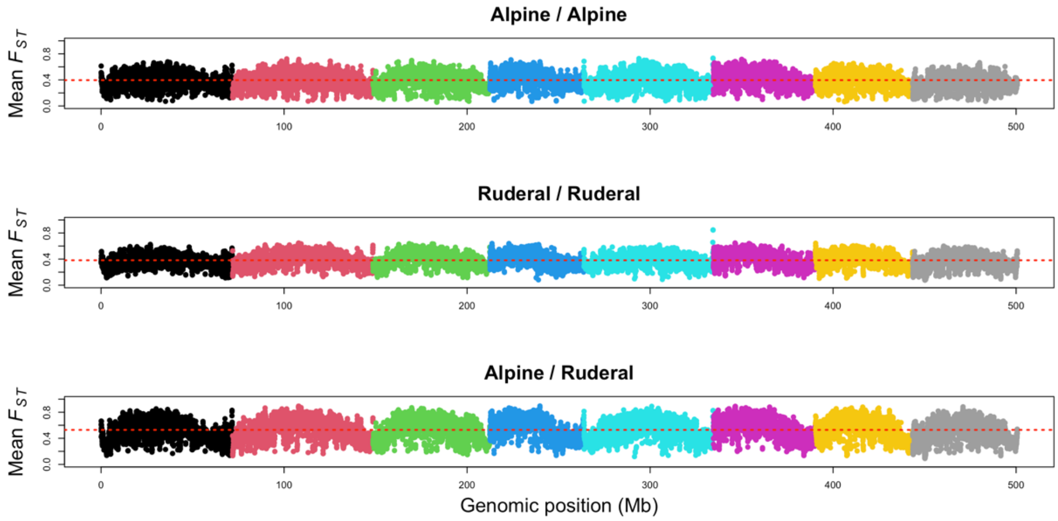


Figure 4.3: Mean F_{ST} across growth habits

Whole genome F_{ST} comparisons between populations, grouped according to growth habit. The mean F_{ST} is represented by a dotted red line. Sequential changes of colour represent different chromosomes (1-8).

To look for islands of relative divergence between alpine and ruderal species, whole genome F_{ST} comparisons were carried out. Figure 4.3 shows the mean F_{ST} from alpine against alpine comparison (a/a), ruderal against ruderal comparisons (r/r), and alpine against ruderal comparisons (a/r). Mean F_{ST} was high in all three cases (mean $F_{ST} > 0.38$). a/r comparisons showed higher mean F_{ST} compared to a/a and r/r comparisons, which showed similar mean F_{ST}. However, the published whole genome phylogeny from Duran-Castillo *et al.* (2021) suggests that *A. molle* and *A. braun-blanquetti* are more genomically similar to species that don't share their growth habit. To observe this genomic similarity, *A. molle* was compared to all ruderal species (except *A. braun-blanquetti*), and *A. braun-blanquetti* was compared to all Kickxiella species (except *A. molle*). As with previous comparisons, distinct peaks were not detectable against the elevated background F_{ST} (Figure 4.4). Mean F_{ST} in both cases was elevated compared to F_{ST}. estimates from a/a and r/r comparisons.

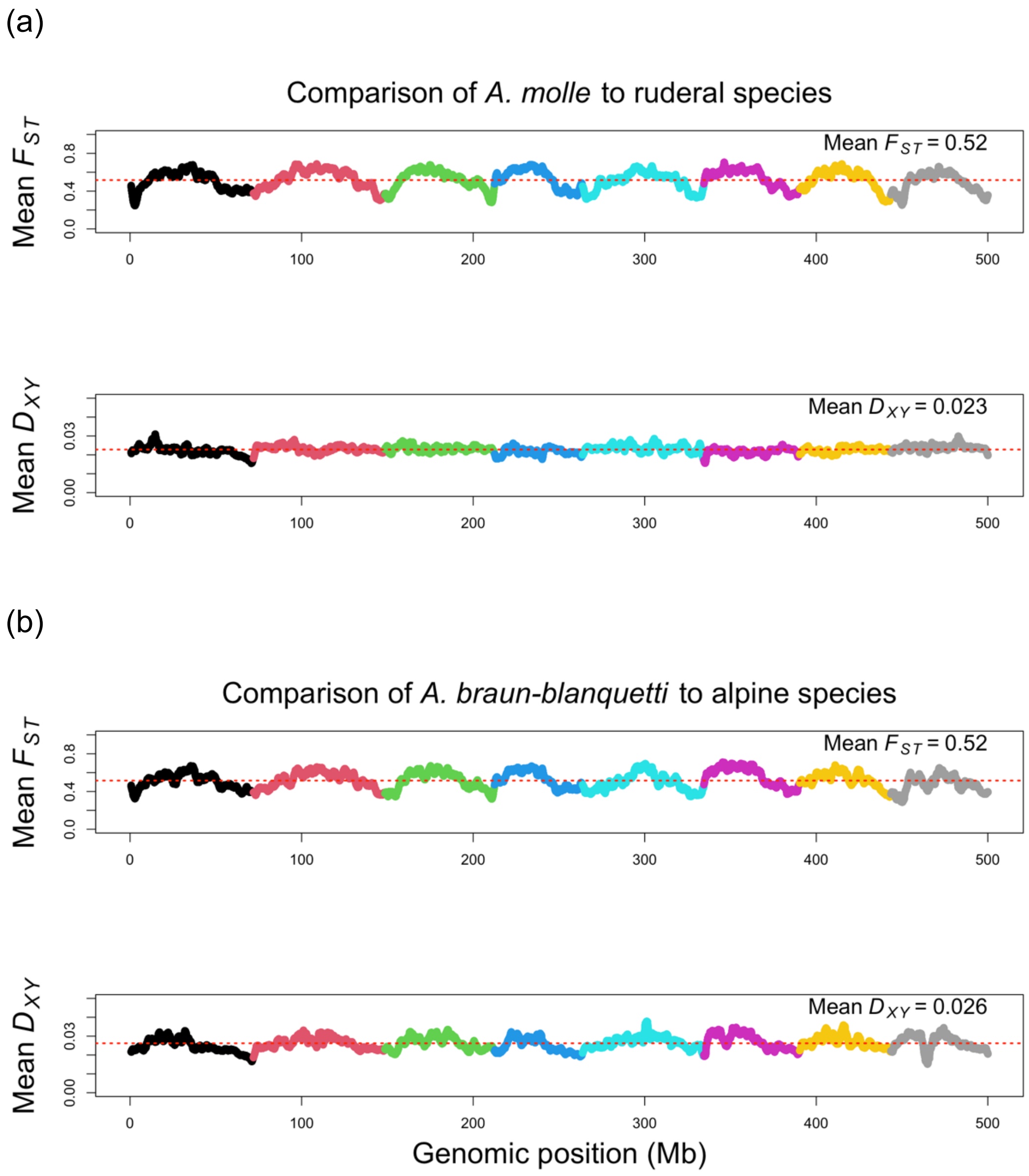


Figure 4.4: Comparisons of A. molle / A. braun-blanquetti to species with different growth habits

(a) Whole genome mean F_{ST} and D_{XY} comparisons between the two *A. molle* pools, and six ruderal pools (excluding *A. braun-blanquetti*). (b) Whole genome F_{ST} and D_{XY} comparisons between the two *A. braun-blanquetti* pools, and six alpine pools (excluding *A. molle*). Dotted red lines indicate the genomic mean. Results have been averaged from 50 kb overlapping windows into 2.5 Mb groups. Distinct colours represent different chromosomes (1-8).

Averaging across all comparisons showed lower F_{ST} values towards the chromosome ends, and elevated values towards the centres (Mean F_{ST} ends = 0.42, Mean F_{ST} centres = 0.50). To investigate how patterns of F_{ST} variation related to within- or between-population diversity, D_{XY} and π_W were averaged across each chromosome in non-overlapping 2.5 Mb windows (Figure 4.5). Mean D_{XY} values were, on average, three times greater than average π_W values. π_W showed a consistent reduction towards the middle of chromosomes, whereas no clear positional effect was apparent for D_{XY}. To quantify this positional effect, chromosomes were split into quarters. The outer two quarters were defined as the chromosome end regions, and the inner two as chromosome centres. Averaging D_{XY} and π_w across the centre and end regions of all chromosomes revealed a reduction in mean π_W (Mean π_W ends = 0.0097, Mean π_W centres = 0.0074), but no significant difference in D_{XY} (Mean D_{XY} ends = 0.023, Mean D_{XY} centres = 0.022). These results suggest that globally elevated F_{ST} is due to high D_{XY} rather than low within population diversity, but patterns of higher and lower F_{ST} values reflect π_W variation.

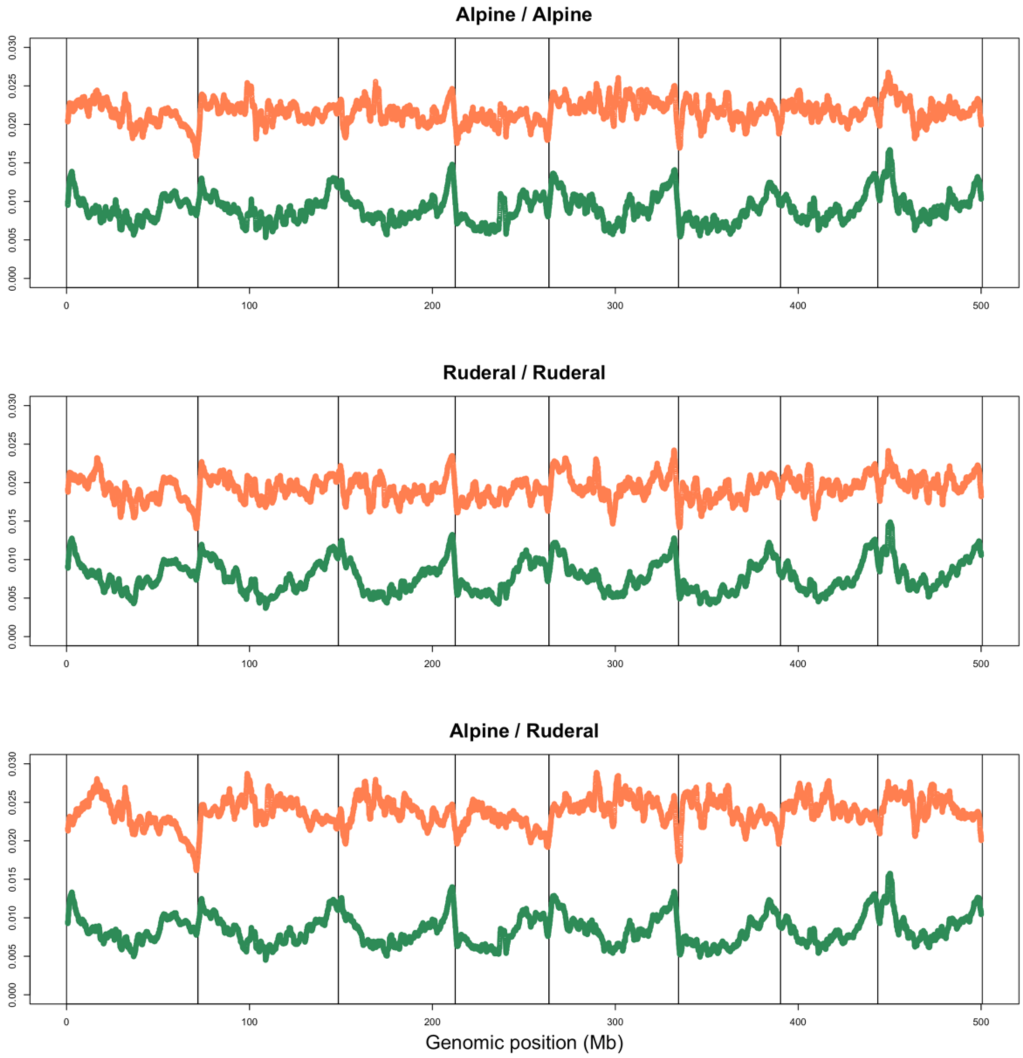


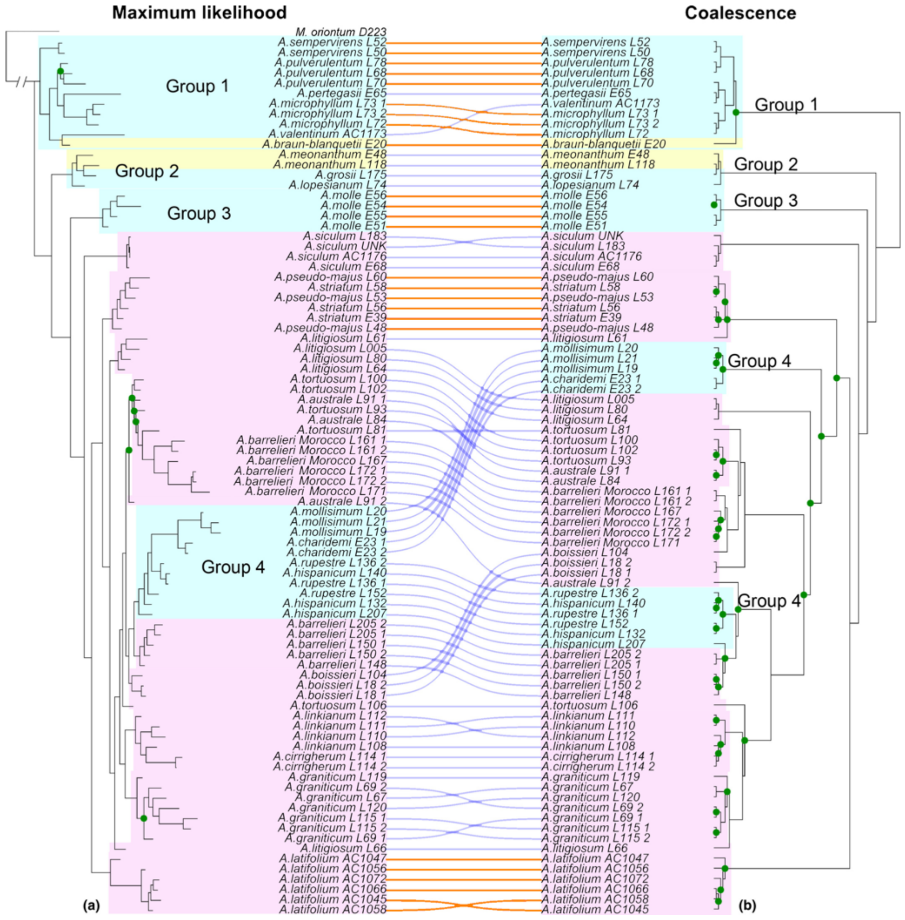
Figure 4.5: D_{XY} and π_{w}, averaged across all alpine, ruderal, and mixed comparisons.

Whole genome mean D_{XY} (orange) and π_{w} (green) averaged across all comparisons between alpine species (top), ruderal species (middle), and mixed alpine and ruderal species (bottom). Values from 50 kb windows have been averaged into 2.5 Mb groups. Black boxes indicate the boundaries of each of the eight chromosomes.

Results: 2 % of subgenomic D_{XY} trees are monophyletic for ruderal and alpine habits

To observe the relationships between these eight species pools based on whole genomic divergence, a whole genome mean D_{XY} phylogeny was constructed (Figure 4.6b). This phylogeny is consistent with results from RAD sequencing (Duran-Castillo *et al.* 2022). Notably, this tree splits up the alpine group, with *A. molle* forming a sister group with the ruderal Antirrhinum group clade. Additionally, the ruderal species *A. braun-blanquetti* formed a sister group with the alpine clade. The placings of these species may be the result of extensive gene flow between alpine and ruderal species, or may reflect parallel evolution of alpine and ruderal habits.

(a)


Maximum likelihood
Coalescence
Group 1
Group 2
Group 3
Group 4
Group 1
Group 2
Group 3
Group 4
Group 4
M. oriontum D223
A.sempervirens L52
A.sempervirens L50
A.pulverulentum L78
A.pulverulentum L68
A.pulverulentum L70
A.pertegasii E65
A.microphyllum L73 1
A.microphyllum L73 2
A.microphyllum L72
A.valentinum AC1173
A.braun-blanquetii E20
A.meonanthum E48
A.meonanthum L118
A.grosii L175
A.lopesianum L74
A.molle E56
A.molle E54
A.molle E55
A.molle E51
A.siculum L183
A.siculum UNK
A.siculum AC1176
A.siculum E68
A.pseudo-majus L60
A.striatum L58
A.pseudo-majus L53
A.striatum L56
A.striatum E39
A.pseudo-majus L48
A.litigiosum L61
A.litigiosum L005
A.litigiosum L80
A.litigiosum L64
A.tortuosum L100
A.tortuosum L102
A.australe L91 1
A.tortuosum L93
A.australe L84
A.tortuosum L81
A.barrelieri Morocco L161 1
A.barrelieri Morocco L161 2
A.barrelieri Morocco L167
A.barrelieri Morocco L172 1
A.barrelieri Morocco L172 2
A.barrelieri Morocco L171
A.australe L91 2
A.mollisimum L20
A.mollisimum L21
A.mollisimum L19
A.charidemi E23 1
A.charidemi E23 2
A.rupestre L136 2
A.hispanicum L140
A.rupestre L136 1
A.rupestre L152
A.hispanicum L132
A.hispanicum L207
A.barrelieri L205 2
A.barrelieri L205 1
A.barrelieri L150 1
A.barrelieri L150 2
A.barrelieri L148
A.boissieri L104
A.boissieri L18 2
A.boissieri L18 1
A.tortuosum L106
A.linkianum L112
A.linkianum L111
A.linkianum L110
A.linkianum L108
A.cirrigherum L114 1
A.cirrigherum L114 2
A.graniticum L119
A.graniticum L69 2
A.graniticum L67
A.graniticum L120
A.graniticum L115 1
A.graniticum L115 2
A.graniticum L69 1
A.litigiosum L66
A.latifolium AC1047
A.latifolium AC1056
A.latifolium AC1072
A.latifolium AC1066
A.latifolium AC1045
A.latifolium AC1058
A.sempervirens L52
A.sempervirens L50
A.pulverulentum L78
A.pulverulentum L68
A.pulverulentum L70
A.pertegasii E65
A.valentinum AC1173
A.microphyllum L73 1
A.microphyllum L73 2
A.microphyllum L72
A.braun-blanquetii E20
A.meonanthum E48
A.meonanthum L118
A.grosii L175
A.lopesianum L74
A.molle E56
A.molle E54
A.molle E55
A.molle E51
A.siculum UNK
A.siculum L183
A.siculum AC1176
A.siculum E68
A.pseudo-majus L60
A.striatum L58
A.pseudo-majus L53
A.striatum L56
A.striatum E39
A.pseudo-majus L48
A.litigiosum L61
A.mollisimum L20
A.mollisimum L21
A.mollisimum L19
A.charidemi E23 1
A.charidemi E23 2
A.litigiosum L005
A.litigiosum L80
A.litigiosum L64
A.tortuosum L81
A.tortuosum L100
A.tortuosum L102
A.tortuosum L93
A.australe L91 1
A.australe L84
A.barrelieri Morocco L161 1
A.barrelieri Morocco L161 2
A.barrelieri Morocco L167
A.barrelieri Morocco L172 1
A.barrelieri Morocco L172 2
A.barrelieri Morocco L171
A.boissieri L104
A.boissieri L18 2
A.boissieri L18 1
A.australe L91 2
A.rupestre L136 2
A.hispanicum L140
A.rupestre L136 1
A.rupestre L152
A.hispanicum L132
A.hispanicum L207
A.barrelieri L205 2
A.barrelieri L205 1
A.barrelieri L150 1
A.barrelieri L150 2
A.barrelieri L148
A.tortuosum L106
A.linkianum L111
A.linkianum L110
A.linkianum L112
A.linkianum L108
A.cirrigherum L114 1
A.cirrigherum L114 2
A.graniticum L119
A.graniticum L67
A.graniticum L120
A.graniticum L69 2
A.graniticum L69 1
A.graniticum L115 1
A.graniticum L115 2
A.litigiosum L66
A.latifolium AC1047
A.latifolium AC1056
A.latifolium AC1072
A.latifolium AC1066
A.latifolium AC1058
A.latifolium AC1045
(a)
(b)

(b)

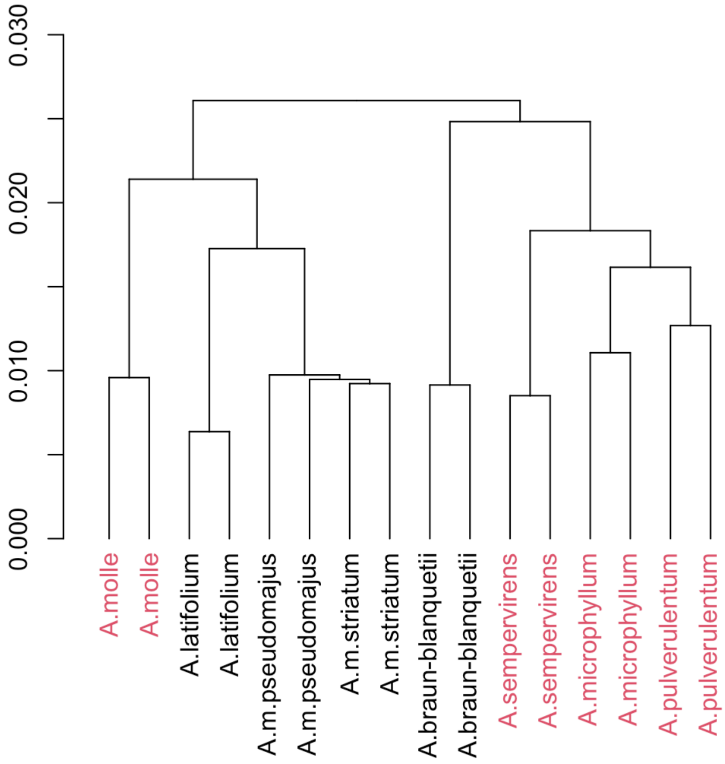
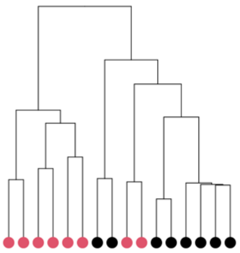
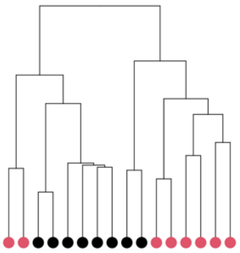
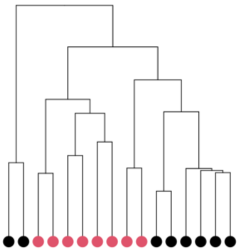
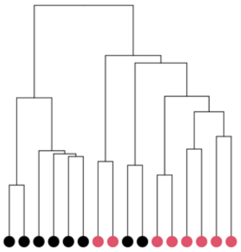
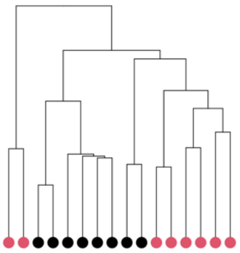

0.030
0.020
0.010
0.000
A.molle
A.molle
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
A.m.striatum
A.braun-blanquetii
A.braun-blanquetii
A.sempervirens
A.sempervirens
A.microphyllum
A.microphyllum
A.pulverulentum
A.pulverulentum

Figure 4.6 (previous page): Whole genome mean D_{XY} tree for 16 species pools

(a) Maximum Likelihood (left) and Coalescence (right) trees generated by Mario Durán-Castillo from RAD-seq data from 28 *Antirrhinum* taxa, with *Misopates orontium* as an outgroup. Lines between the two trees indicate the relative positions of taxa on both trees – where this line is orange, the adjacent species is also present within Figure 4.6b, and proceeding analyses within this chapter. Reproduced and adapted with permission from Durán-Castillo *et al.*, 2022. (b) A D_{XY} tree summarising the relationships between the 16 experimental populations. Mean D_{XY} between each pair of populations was averaged over all 19,502 50 kb genomic windows. Taxa in red have an alpine habit, and black taxa have a ruderal habit. Observed groupings are consistent with those in Figure 4.6a.

To explore how patterns of divergence vary across the genome, subgenomic D_{XY} trees were constructed. Mean D_{XY} was sampled across the whole genome in 50 kb windows, with a 25 kb overlap, yielding 19,502 windows. Subgenomic trees were constructed from D_{XY} distance matrices using UPGMA. To classify topologies, each tree was split at its root division, yielding two clades. In doing this, a simple representation of population grouping can be derived for each tree. For example, splitting tree that is monophyletic for both growth habits would yield two groups, one containing four alpine species and the other containing four ruderal species. This can be compactly represented as aaaa:rrrr. The whole genome tree (Figure 4.6) can be represented as arrr:aaar. A total of 275 distinct tree topologies were detected, based on root division. The 10 most frequent subgenomic tree topologies are shown in Table 4.1. Of particular interest in studying growth habit divergence are trees which are monophyletic for both ruderal and alpine growth habits (aaaa:rrrr). 2 % of all genomic trees (427 / 19,502 trees) gave doubly monophyletic growth habit groupings (Figure 4.7). 45 % of doubly monophyletic trees had elevated SRB compared to the genomic mean. Deeply rooted, doubly monophyletic trees represent the strongest signatures of differential growth habit adaptation. The *Hairy* region tree was within the top 4 % of all doubly monophyletic trees based on SRB, confirming *Hairy* as a candidate gene involved in growth habit divergence. The 427 identified windows giving monophyletic trees will be termed monophyletic regions.

Rank	*Percentage of subgenomic trees*	*Topology code*	*Mean Topology*	*Description*
1	*23.4 %*	`aaa:arrrr`		Polyphyletic for both growth habits
2	*21.2 %*	`arrr:aaar`		Polyphyletic for both growth habits. Identical to the mean whole genomic topology.
3	*12.1 %*	`r:aaaarrr`		Polyphyletic for both growth habits
4	*6.5 %*	`rrr:aaaar`		Polyphyletic for both growth habits
5	*6.2 %*	`a:aaarrrr`		Polyphyletic for both growth habits

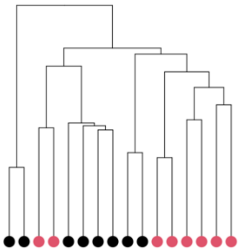
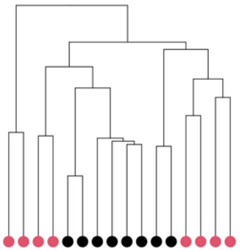
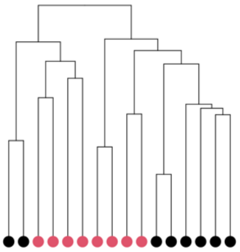
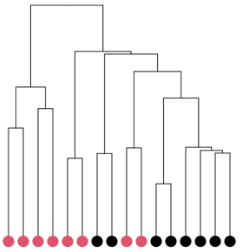
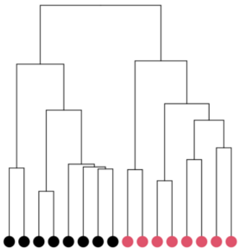
6	*4.0 %*	`r:aaaarrr`		Polyphyletic for both growth habits. *A. latifolium* outlying.
7	*3.1 %*	`a:aaarrrr`		Polyphyletic for both growth habits. *A. sempervirens* outlying.
8	*2.9 %*	`aar:aarrr`		Polyphyletic for both growth habits
9	*2.6 %*	`aa:aarrrr`		Polyphyletic for both growth habits
10	*2.2 %*	`aaaa:rrrr`		Monophyletic for both growth habits

Table 4.1: The top 10 most common tree topologies.

Genomic frequencies, compact representations, topologies, and descriptions of the six distinct trees represented within the 54 consistent outlier windows. Within topology codes, *a* refers to an alpine population, and *r* to a ruderal population. On the trees, alpine taxa are represented by red circles, and ruderal taxa by black circles.

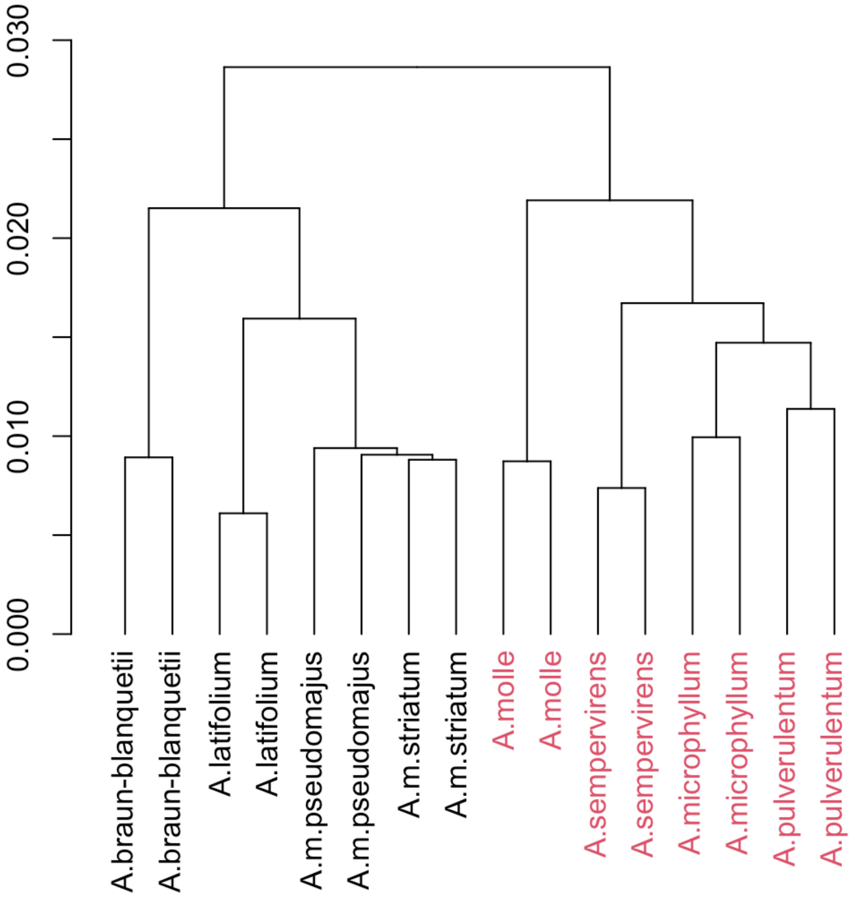


Figure 4.7: Mean D_{XY} tree for monophyletic regions

Mean D_{XY} tree topology obtained from averaging all D_{XY} comparison results across the 427 subgenomic regions yielding doubly monophyletic trees. Taxa in red have an alpine habit, and black taxa have a ruderal habit.

The top 10 subgenomic topologies accounted for 84 % of all trees. The doubly monophyletic class described above was the tenth most common genomic topology. No other topologies in the top ten were monophyletic for either habit. The most common topology (aaa:arrrr) occurred 4,555 times (23.4 % of all subgenomic trees). Like the whole genome tree, it gave groupings that were polyphyletic for both growth habits (Figure 4.8). This polyphyly was due to one species, *A. molle*, being nested within a clade of otherwise ruderal species. The second most common topology (arrr:aaar) was identical to the whole genome tree. Four topologies reflected cases where a single species was an outlier, most commonly *A. braun-blanquetti* (12.5 % of subgenomic trees). The remaining 16 % of trees showed 217 diverse topologies.

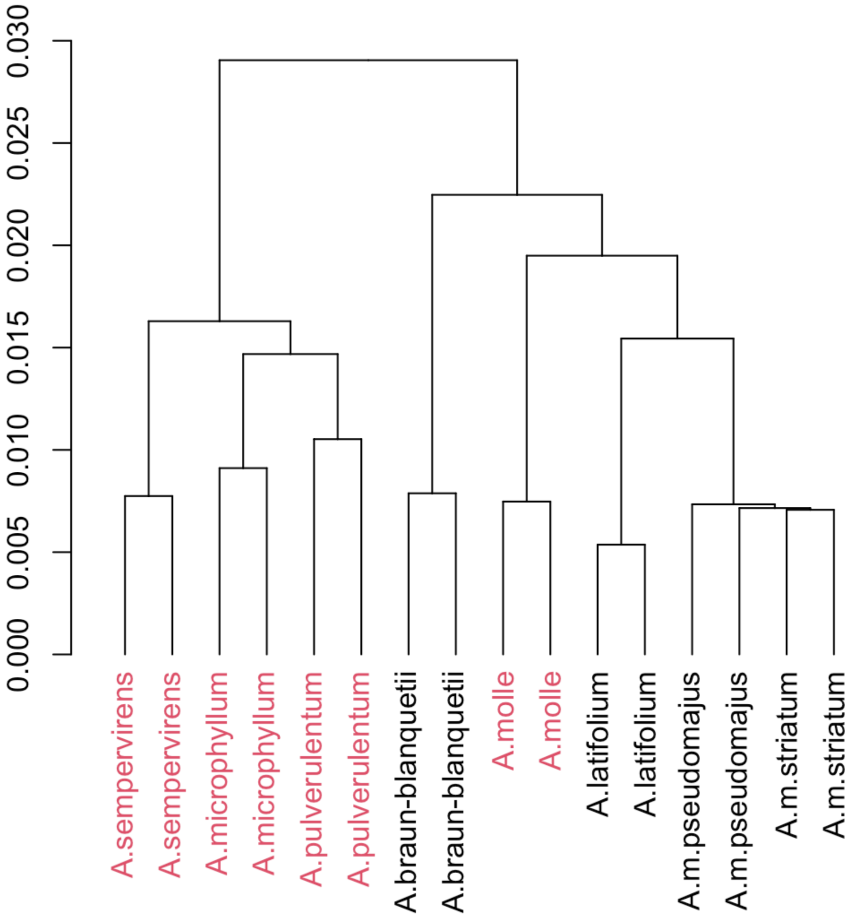


Figure 4.8: Mean D_{XY} tree of the most common subgenomic topology

Mean D_{XY} tree topology from the 4,555 trees showing the most common subgenomic topology. Taxa in red have an alpine habit, and black taxa have a ruderal habit.

Results: Monophyletic islands span at least 7 kb on average, and show elevated D_{XY}

To distinguish monophyletic regions that are enriched for SNPs showing monophyletic distributions, D_{XY} trees were generated in 1 kb windows, with 900 bp overlaps, across all 427 identified 50 kb monophyletic regions. Many regions had a significant amount of missing data at this increased resolution, consistent with low sequencing depth. Low-depth monophyletic regions, returning < 100 / 499 possible 1 kb window trees, were excluded, leaving 394 regions. 1 kb window trees from each remaining region were then classified according to whether they gave monophyletic (aaaa:rrrr) or non-monophyletic groupings, based on root division. The number of 1 kb windows giving monophyletic trees, and the total number of trees recovered, was recorded for each region (trees could not be constructed where read depth was low in one or more populations). The greatest reported proportion of monophyletic trees within a 50 kb region was 0.98 (378 / 387 1 kb window trees being monophyletic). Given that 1 kb windows have 900 bp overlaps, each adjacent window covers an additional 100 bp of sequence. To estimate the sizes of monophyletic islands across all monophyletic regions, minimum island size was estimated. By making the conservative assumption that all 1 kb windows were directly adjacent, I could calculate minimum island size as the total number of monophyletic 1 kb trees multiplied by the 100 bp interval. The largest monophyletic island observed spanned at least 37.8 kb. This is more than 126 times greater than predicted by the parallel-allele-recruitment hypothesis, but consistent with barriers to gene flow or allele sharing. Minimum island size was calculated for all 427 50 kb monophyletic regions (Figure 4.9). The mean minimum island size was 7.1 kb, demonstrating that monophyletic islands are generally significantly larger than expected by the parallel-allele-recruitment hypothesis. Eight islands had a minimum size of zero, reflecting the complete absence of 1 kb window monophyletic trees. Here, 50 kb monophyletic trees are likely artifacts of averaging many trees with different topologies.

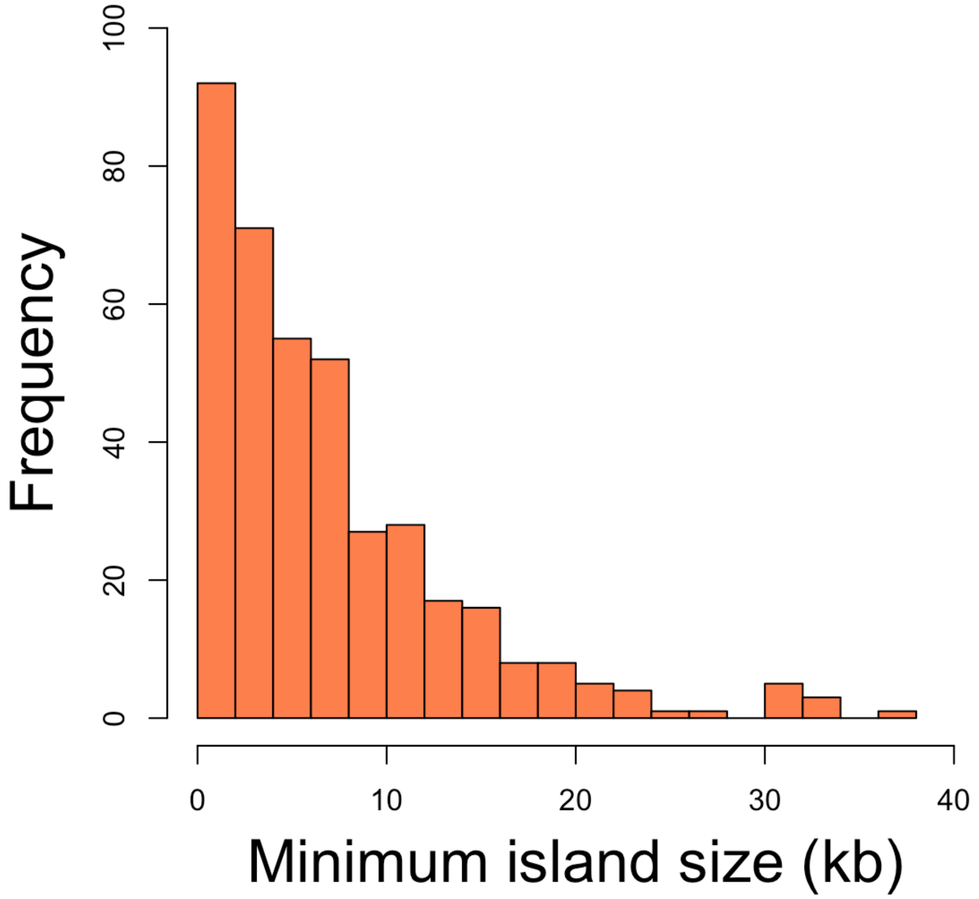


Figure 4.9: Estimates of minimum monophyletic island size

Histogram of minimum monophyletic island size from the 427 monophyletic regions.

Monophyletic regions were slightly enriched within the central regions of chromosomes (59 %) compared to chromosome ends (41 %). Islands within central chromosome regions were larger (mean minimum size = 7.7 kb) than those within outer regions (mean minimum size = 6.1 kb).

To quantify the extent to which monophyletic islands extend beyond 50 kb boundaries, a subset of monophyletic islands were reanalysed with 500 kb flanks included. These islands were selected based on the *SRB* of their 50 kb trees. Islands with the five greatest and five smallest *SRB* values were selected, along with five showing intermediate values

Three plots have been included within this chapter (Figures 4.10 – 4.12), showing one region from each of these *SRB* classifications. The remaining plots are shown in Appendix 5. Where multiple islands overlapped, the islands with lower *SRB* values were excluded. Island size correlated with *SRB*, with deeply rooted 50 kb trees reflecting a greater mean monophyletic island size on average (Table 4.2). However, the largest island observed did not have a large *SRB* value, indicating that a deeply rooted tree is not necessarily reflective of a larger monophyletic island. The 15 surveyed monophyletic islands showed elevated D_{XY}, and regions of very low π_W. To test whether this was a general trend across monophyletic islands, mean diversity statistics were calculated from 50 kb window data (Figure 4.13). Monophyletic islands showed significantly elevated D_{XY} in comparisons between alpine and ruderal species (t-test, $p < 2.2 \times 10^{16}$). Significant differences in π_W were not observed for alpine or ruderal species at monophyletic islands (t-test, $p = 0.14$).

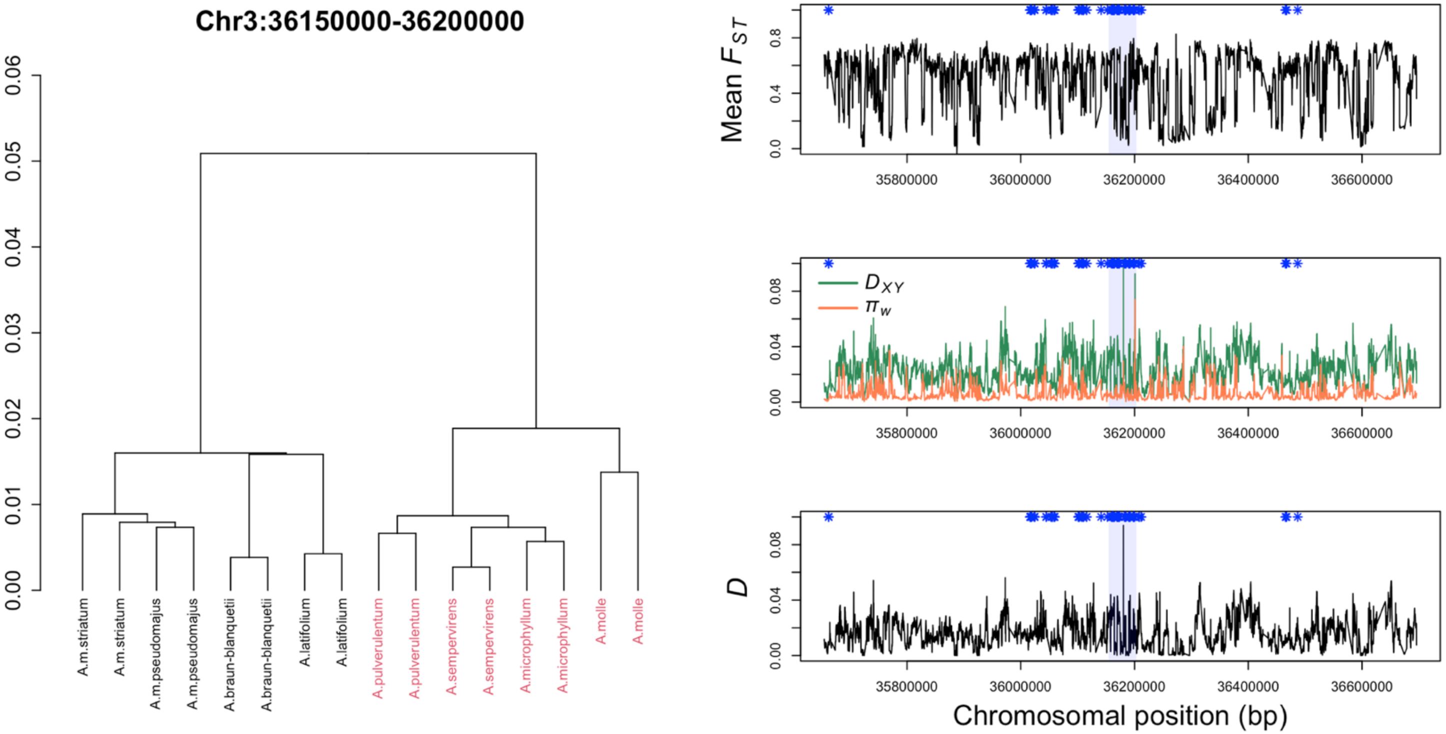


Figure 4.10: Plots summarising a monophyletic region showing high SRB

(left) Summarising 50 kb window tree for the whole monophyletic region. (right, top) Mean F_{ST} across the monophyletic region from comparisons of all populations, calculated in 1 kb windows with 900 bp overlaps. (right, middle) Mean D_{XY} (green) from all population comparisons, and mean π_w (orange) from all populations, summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. (right, bottom) Nei's D (D_{XY} - π_w) summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. Blue asterisks indicate the locations of 1 kb window trees that are monophyletic for alpine and ruderal growth habits. The pale blue area is the originating 50 kb monophyletic region.

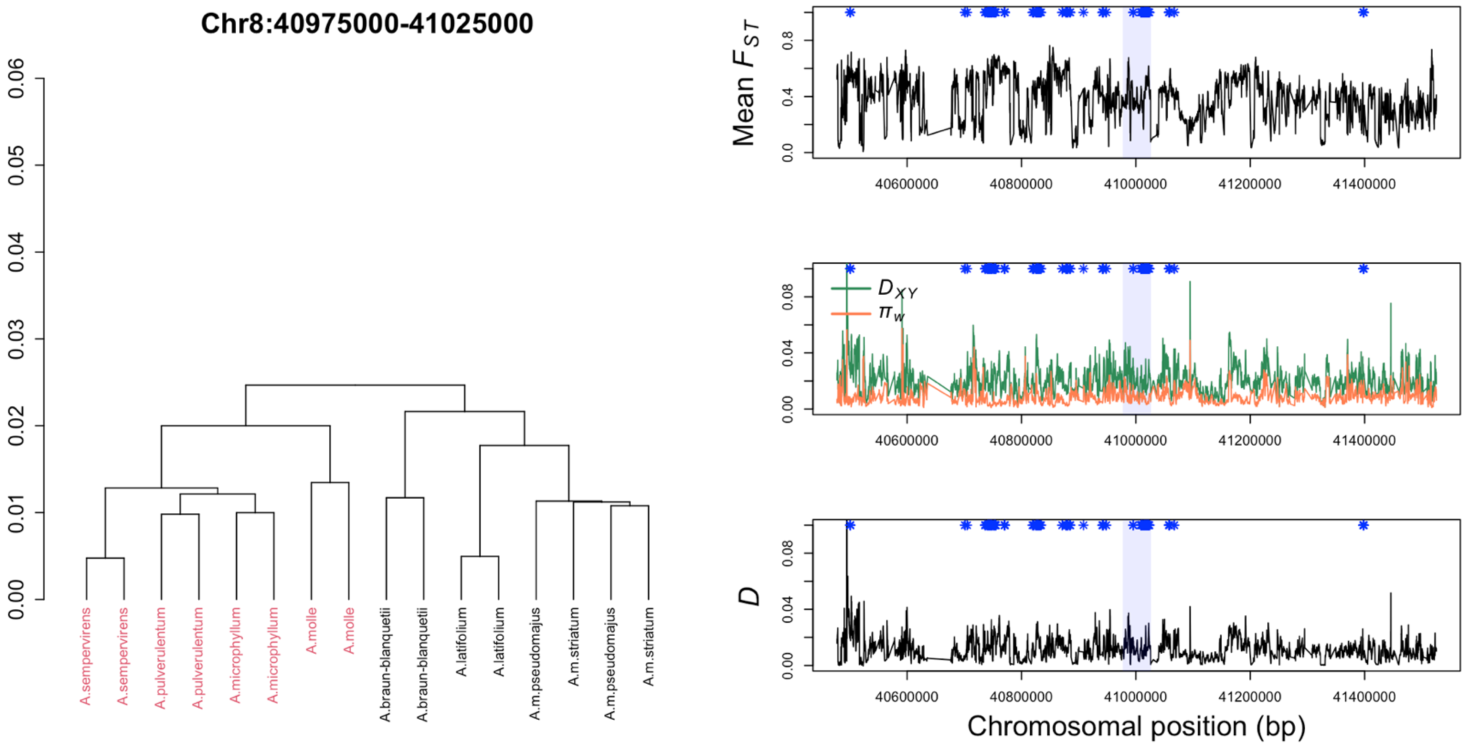


Figure 4.11: Plots summarising a monophyletic region showing intermediate SRB

(left) Summarising 50 kb window tree for the whole monophyletic region. (right, top) Mean F_{ST} across the monophyletic region from comparisons of all populations, calculated in 1 kb windows with 900 bp overlaps. (right, middle) Mean D_{XY} (green) from all population comparisons, and mean π_w (orange) from all populations, summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. (right, bottom) Nei's $D\,(D_{XY}$ - $\pi_w)$ summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. Blue asterisks indicate the locations of 1 kb window trees that are monophyletic for alpine and ruderal growth habits. The pale blue area is the originating 50 kb monophyletic region.

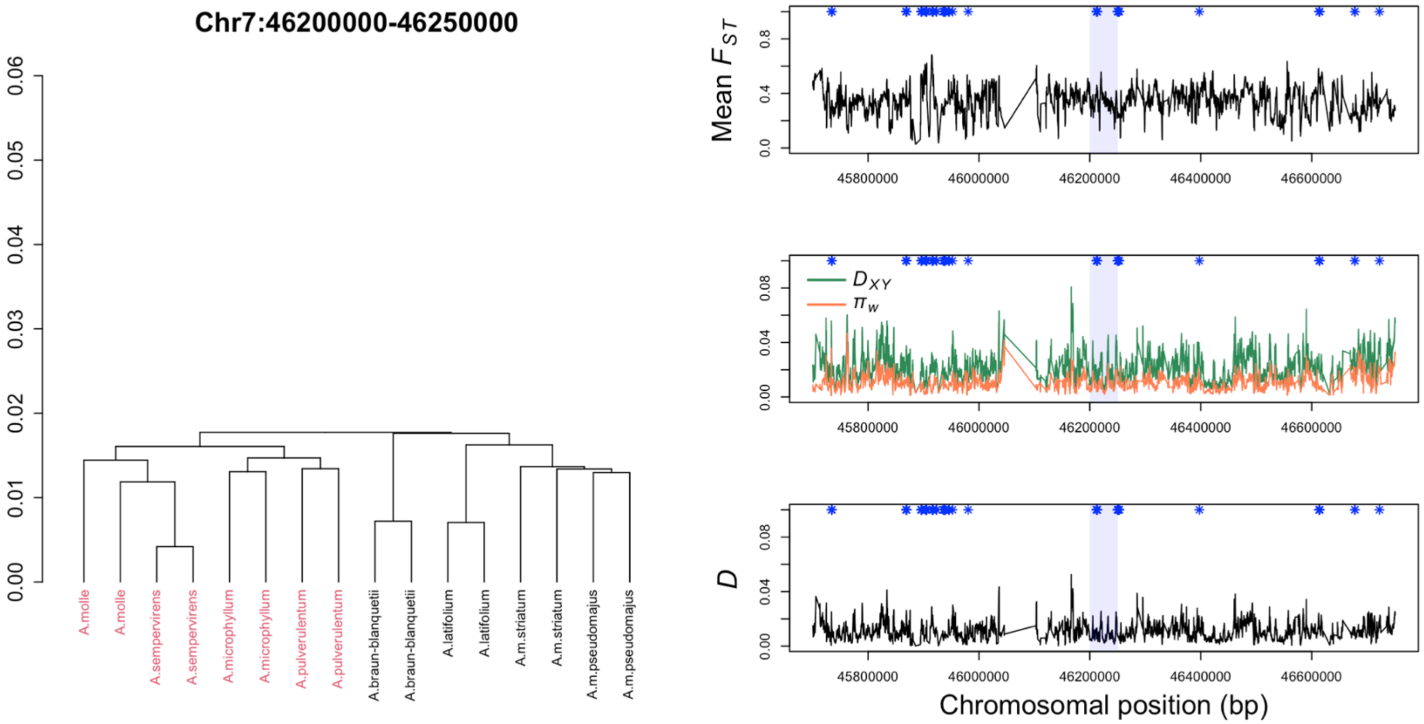


Figure 4.12: Plots summarising a monophyletic region showing low SRB

(left) Summarising 50 kb window tree for the whole monophyletic region. (right, top) Mean F_{ST} across the monophyletic region from comparisons of all populations, calculated in 1 kb windows with 900 bp overlaps. (right, middle) Mean D_{XY} (green) from all population comparisons, and mean π_w (orange) from all populations, summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. (right, bottom) Nei's D (D_{XY} - π_w) summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. Blue asterisks indicate the locations of 1 kb window trees that are monophyletic for alpine and ruderal growth habits. The pale blue area is the originating 50 kb monophyletic region.

Monophyletic region	**50 kb tree *SRB***	**Estimated monophyletic island size (bp)**
Chr3:36150000-36200000	0.0320	200,000
Chr3:51050000-51100000	0.0244	50,000
Chr6:23650000-23700000	0.0225	60,000
Chr7:34350000-34400000	0.0242	150,000
Chr8:27475000-27525000	0.0317	250,000
Chr2:18600000-18650000	0.00311	10,000
Chr3:57550000-57600000	0.00301	10,000
Chr5:18675000-18725000	0.00303	10,000
Chr5:56175000-56225000	0.00306	40,000
Chr8:40975000-41025000	0.00303	375,000
Chr1:24675000-24725000	0.000155	70,000
Chr4:6750000-6800000	0.000107	60,000
Chr6:49950000-50000000	0.000154	0
Chr7:46200000-46250000	0.000114	1,000
Chr8:49175000-49225000	0.000116	1,000

Table 4.2: SRB and monophyletic island size of 15 monophyletic regions

Estimated monophyletic island sizes for islands showing the five largest *SRB* values (green background), the five median *SRB* values (yellow background), and five lowest *SRB* values (orange background).

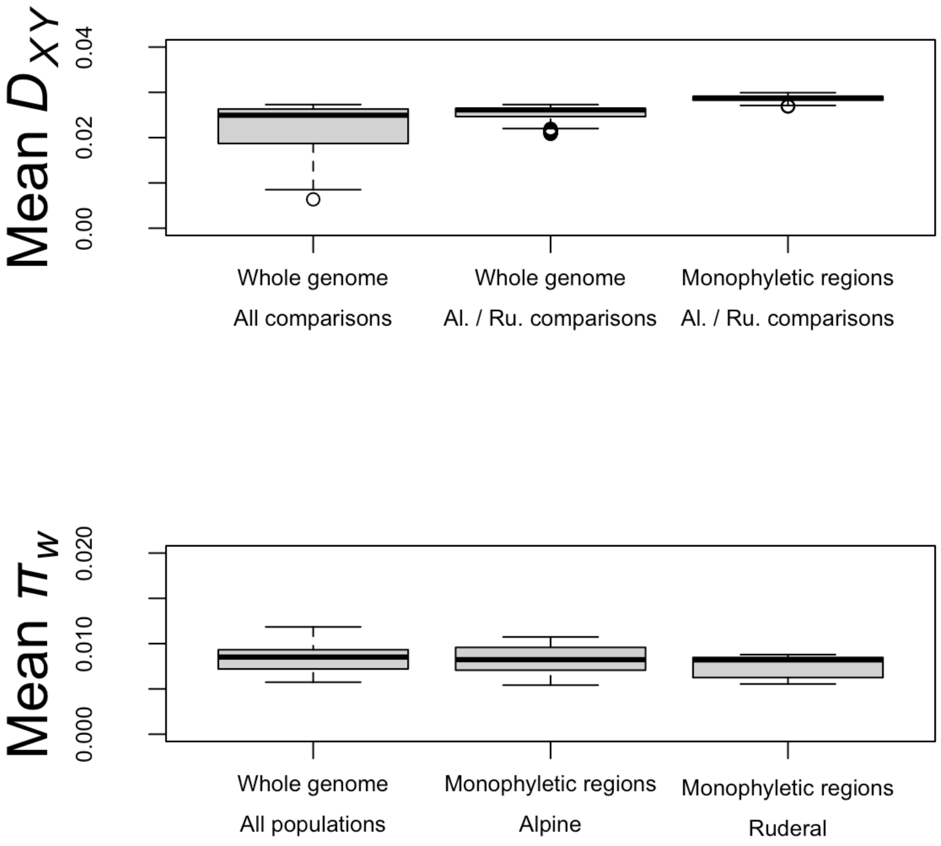


Figure 4.13: Boxplots summarising mean D_{XY} and π_w across the whole genome, and monophyletic regions.

Boxplots summarising mean D_{XY} (top) and π_W (bottom) across the whole genome, compared to windows within monophyletic regions. *Al* and *Ru* refer to alpine and ruderal populations respectively. Mean D_{XY} is significantly higher within monophyletic regions than the whole genomic average calculated from all comparisons, and from only alpine and ruderal comparisons (t-test, $p < 2.2\times10^{-16}$ in both cases).

Discussion

Monophyletic signatures pinpoint genomic islands of divergence around growth habit

This chapter aimed to identify genomic regions segregating between *Antirrhinum* species showing different growth habits, in order to test hypotheses about the evolutionary dynamics of growth habit divergence. This has been achieved by adapting the grouping tree scan methodology to search for a specific monophyletic signature dividing populations based on growth habit. A total of 427 monophyletic regions were identified, suggesting that the adaptive architecture of growth habit is significantly more complicated than flower colour. However, for a polygenic trait, this is actually a surprisingly small number of loci, particularly as several are likely to be artifacts. Until monophyletic islands are further characterised, it is difficult to conclude whether they are likely to be involved in growth habit divergence, or whether they contain genes of strong or weak adaptive effects. However, by considering the properties of the identified islands, hypotheses pertaining to their origins can be tested.

Monophyletic islands reflect barrier genes, or allele sharing

Growth habit is a multifaceted trait that is likely to have a complex genetic architecture. Therefore, if distinct growth habits arose through parallel *de novo* mutations, the genetic architecture is unlikely to be consistent between different species. Analysis of an F_2 population generated from crossing *A. molle* and *A. sempervirens* showed evidence of segregation of major growth habit traits, suggesting that the genetic basis of growth habit divergence is shared between both species (Li *et al.*, Preliminary data). This direct evidence of shared variation affirms the idea that growth habit divergence likely involved some degree of allele sharing. The uptake of ancient alleles through introgressive hybridisation is regarded as an important driver of adaptive evolution (Marques, Meier, and Seehausen, 2019). Results presented here suggest that, even using the inherently conservative minimum island size approach, monophyletic islands are generally significantly larger than predicted by the parallel allele recruitment hypothesis. This argues that allele sharing through hybridisation is a more likely scenario in the case of convergent growth habit evolution.

The hypothesised parallel evolution of the alpine growth habit from a ruderal ancestral state is based on findings from whole genome phylogenetic analyses (Wilson and Hudson, 2011, Duran-Castillo *et al.*, 2022). However, whole genomic trees cannot adequately summarise the heterogenous nature of genomic divergence, making it important to consider other hypotheses. In the simplest scenario, the divergence of the alpine and ruderal growth habits occurred in a single event. In this case, monophyletic islands may contain longstanding barriers to gene flow involved in ancient divergence. The grouping of the alpine *A. molle* with ruderal species within the whole genome tree may reflect extensive gene flow across non adaptive loci. Indeed, putative hybrids of *A. molle* and *A. m. pseudomajus* have been observed in the field (Coen, unpublished data). The existence of a major subgenomic tree topology grouping *A. braun-blanquetti* within the ruderal clade may reflect shared ancestral alleles between *A. braun-blanquetti* and other ruderal species, which contradicts the postulated parallel evolution of the Streptosepalum and Antirrhinum groups. Alternatively, *A. braun-blanquetti* may have also engaged in gene flow with species in the Antirrhinum group, as predicted by ABBA-BABA analysis (Duran-Castillo et al., 2022).

To differentiate between ancient barrier genes and allele sharing, future work could test how consistently growth habit islands occur within different populations. The allele sharing hypothesis proposes that variation arose across a range of populations, and was shared through secondary contact. Due to the unlikeliness of all populations having made contact, it is expected that the distribution of some growth habit adaptations may be sporadic. In contrast, barrier genes are expected to have arisen during an ancient divergence of alpine and ruderal species, and are therefore expected to be represented across all populations. Preliminary data suggests that certain growth habit traits are sporadic in the wild. Work by Tingting Li, under my supervision, has shown that at least one population of the ruderal species *Antirrhinum latifolium* has a recessive *hairy* allele (Li *et al.*, preliminary data). This was also reported by Tan *et al.* (2020). Analyses presented here only contain alpine species from what Tan *et al.* refer to as the basal Kickxiella group. Species such as *Antirrhinum hispanicum* and *Antirrhinum charidemi* are phenotypically distinct from basal Kickxiella, showing evidence of alpine and ruderal characters (Sutton, 1988). Testing for genomic islands underlying these 'semi-alpine' habits may shed more light onto how growth habits diverge.

Further analyses should explore the gene landscape of monophyletic islands

The detection of monophyletic islands strongly implies genomic regions are involved in growth habit divergence. To test whether monophyletic islands contained genes involved in growth habit, predicted genes (from Li *et al*., 2019) were extracted from all 427 50 kb monophyletic islands. This yielded 1,219 genes. 400 / 427 monophyletic islands contained at least one gene. Predicted genes were then functionally annotated using eggNOG-mapper (Cantalapiedra *et al.*, 2021). A total of 137 genes were annotated. 43 genes were predicted to encode retroviral and transposon associated domains. These genes were excluded, leaving 89 genes split between 83 regions. Many genes were predicted to encode transcription factors, including those with MADS-box and homeobox domains, and those within the AP2/EREBP family. Such transcription factors are strong candidates for regulation of growth (Ramachandran, Hiratsuka, and Chua, 1994, Kaufman and Airoldi, 2018), but functional relevance cannot be implied in the absence of expression data. Differential expression analysis of RNAseq data from a range of tissues may facilitate the discrimination of candidate genes. An advantage of *Antirrhinum* as a model system is that species showing distinct growth habits are inter-fertile. Therefore, the role of identified genes in growth habit divergence can be directly tested by generating hybrid populations showing allelic segregation across panels of candidate genes.

5: *Many SULF*-like, sRNA-producing inverted repeats show presence / absence relationships between *Antirrhinum* species, and may underlie phenotypic divergence

Introduction

Response to selective pressures can drive rapid evolution of traits, particularly when involved in reproductive success (Franks, Sim, and Weis, 2007, Lankinen and Green, 2015, Mackin *et al.*, 2021). Previous chapters have established how colour genes can interact to affect fitness. It is likely therefore that evolution of phenotypic novelty in colour, and similar traits, depends not only on the evolution of novel factors, but also the fine control of existing genes. Flower colour patterns in *Antirrhinum* species involve magenta and yellow pigments. The major yellow pigment in *Antirrhinum* is aurone. Aurone biosynthesis depends on the *FLA* gene on chromosome two (Bradley *et al.*, manuscript in preparation). *FLA* encodes a chalcone 4’-*O*-glucosyltransferase which catalyses the glucosylation of chalcone to aurone, in a biosynthesis pathway with only one intermediate stage (Ono *et al.*, 2006). Biosynthesis of aurone is restricted to only a few taxa in snapdragons (Ellis and Field, 2016), which suggests that the ability to synthesise yellow is a recently acquired trait. Another locus, *SULFUREA* (*SULF*), has emerged as a regulator of yellow patterning. *SULF* is localised to chromosome 4 and acts to repress the spread of yellow by targeting *FLA* transcripts for degradation. The interaction of *SULF* and *FLA* gives rise to the restricted yellow observed in *A. m. pseudomajus*.

Origin and mechanism of *SULF*

In 2017, Bradley *et al.* demonstrated that *SULF* affects *FLA* expression through the generation of regulatory small RNAs (sRNAs). Use of the term sRNA in the broad sense typically refers to any of a range of RNA molecules that act to regulate cellular processes through interactions with other RNAs. It is generally accepted that sRNA silencing originated as a defence mechanism against transposable elements and RNA viruses and has since been adapted to fulfil other functions (Borges and Martienssen, 2015, Eamens *et al.*, 2008). Although diverse in their biogenesis and processing, all sRNAs share the property of generating mature RNAs averaging 20 – 25 nt in length (Hannon *et al.*, 2006, Borges and Martienssen, 2015, Morgado and Johannes, 2017). The length of mature sRNA-derived RNAs varies according to their function. For example, 24 nt sRNAs are mainly associated with gene silencing at the transcriptional level, through RNA-directed DNA methylation (Lewsey *et al.*, 2016). 21 nt sRNAs are also involved in gene silencing, but act post-transcriptionally, by directing the cleavage of complementary mRNAs (Hamilton and Baulcombe, 1999). sRNAs produced by *SULF* are mainly 21 nt (Bradley *et al.*, 2017).

SULF is a hairpin sRNA, originating from an inverted repeat (IR) motif. The *SULF* IR contains regions of inverted sequence homology that can fold back upon themselves, forming a unimolecular double stranded RNA. This double-stranded conformation of immature sRNAs increases their stability, and allows them to escape cellular degradation mechanisms. It also facilitates their recognition by processing factors. In plants, the cleavage of precursor sRNAs to generate mature transcripts is carried out by DICER-LIKE proteins, which recognise double-stranded RNA duplexes (Borges and Martienssen, 2015). Processed transcripts associate with a range of ARGONAUTE proteins, which guide sRNAs to fulfil specific roles depending on their size and sequence specificity, in tandem with other silencing factors (Ma and Zhang, 2018). Many hairpin sRNAs, including *SULF*, have a region of intervening sequence between the two regions of inverted homology. While this sequence may affect the stability of the sRNA molecule, it is not expected to yield mature sRNAs. For brevity, when discussing hairpin sRNAs / inverted repeats, I will refer to the regions of inverted sequence homology as the arms, and the intervening sequence as the spacer (Figure 5.1).

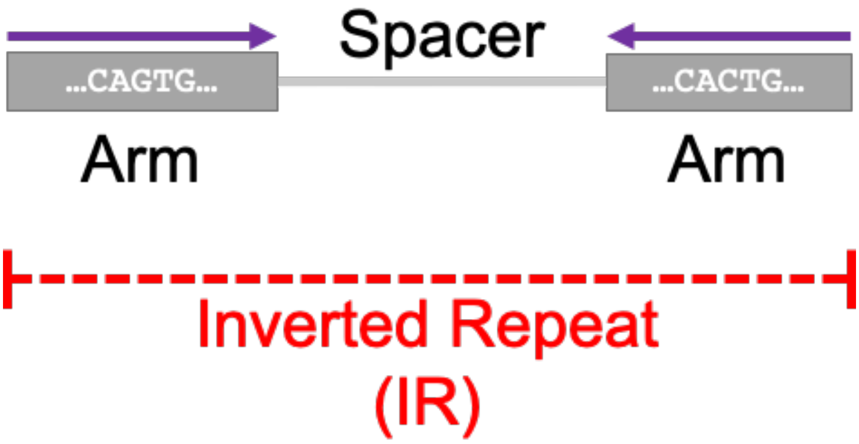


Figure 5.1: Diagrammatic representation of an inverted repeat

The two complementary repeat arms (grey boxes) are separated by a spacer region. Purple arrows represent the direction of the repeat, relative to a hypothetical ancestral paralogue. The dotted red line represents the full span of the IR.

For a mature sRNA to target the transcripts of a specific gene, it must have sequence complementarity to that gene. IRs encoding sRNAs can arise through multiple gene duplication events, where one copy comes to exist in an inverted conformation to the other (Allen *et al*., 2004). *SULF* appears to have arisen from multiple inverted duplications of *FLA* (Figure 5.2). Other sources of sRNA loci include spontaneous evolution from pre-existing genomic IRs, or transmission by IR-containing transposable elements (Cui, You, and Chen, 2017).

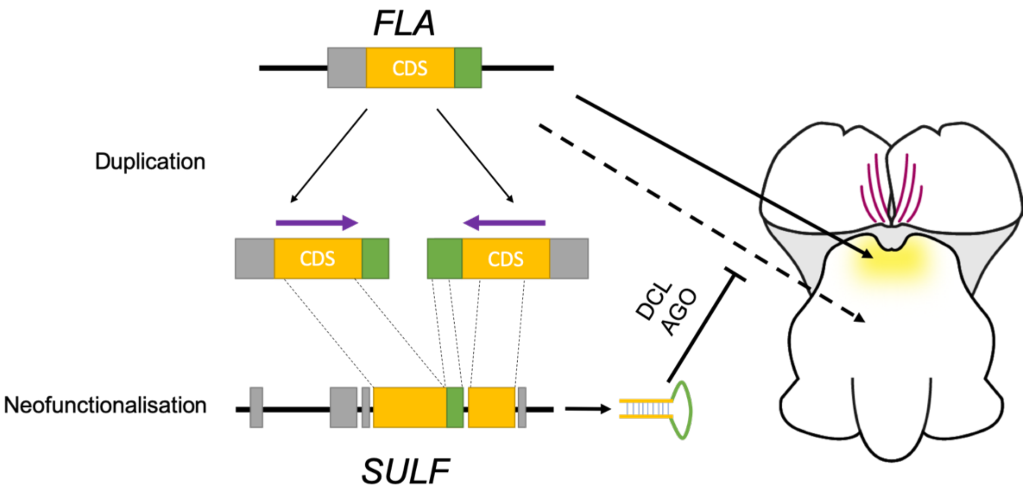


Figure 5.2: Proposed origin of SULF through duplication of FLA.

A simplified schematic showing the hypothesised origin of *SULF*. An inverted duplication of the *FLA* paralogue gives rise to two paralogous copies within close proximity on chromosome 4. The inverted configuration of the gene copies means that they show sequence complementarity. Over time, sequence similarity to the ancestral paralogue is lost, and neofunctionalisation imparts sRNA-producing functionality. Expression of this neofunctionalised locus generates transcripts that are able to form hairpin structures. These are processed by DICER-LIKE (DCL) and ARGONAUTE (AGO) proteins to target *FLA* transcripts for degradation, and thereby locally repress the spread of yellow pigment. Purple arrows indicate the relative orientations of the two *FLA* paralogues. Adapted from Bradley *et al.*, 2017. Snapdragon flower illustration by Mabon Elis.

Functional similarities to microRNAs

Similar to hairpin sRNAs are microRNAs (miRNAs). Like hairpin sRNAs, miRNAs originate from genomic IRs, and are processed to yield sRNAs that are 20-22 nt (Lee and Carroll, 2018). However, miRNA precursor IRs tend to be smaller than hairpin sRNA precursors, and processed transcripts show lower complexity than sRNAs (Morgado and Johannes, 2017). In contrast to metazoan miRNAs, plant miRNAs tend to have specific target sites of high sequence similarity, meaning that they target only a limited number of mRNAs. Both hairpin sRNAs and miRNAs have been linked to phenotypic variation in natural and domesticated plant populations (Debernardi *et al.*, 2017, Clop *et al.*, 2006, Zhang *et al.*, 2018). In each of these examples, variation in function involves allelic variation in known miRNAs, or their targets. The activity of *SULF* is different, in that it depends on the presence of its precursor hairpin. Despite the large size of the *SULF* IR, *SULF* shows low average complexity in its transcripts, akin to a miRNA. It is hypothesised that *SULF* represents an intermediate stage on the pathway to miRNA evolution, and that the high relative sequence similarity to its target gene is a consequence of a recent evolutionary origin. It follows that sequence similarity will wane over time through mutation and drift, and the size of the *SULF* IR will decrease. This has been observed in miRNAs identified within a wide variety of model plant species (Cui, You, and Chen, 2007, Nozawa, Miura, and Nei, 2012).

sRNA-mediated neofunctionalisation

The discovery of *SULF* has provided insights into how the spread of yellow pigmentation is restricted within *A. m. pseudomajus*, but it has also raised questions regarding its origin. It is estimated that around 65 % of plant genes are paralogous, sharing descent with at least one other ancestral gene (Moore and Purugganan, 2005, Panchy, Lehti-Shiu, and Shiu, 2016). When a gene undergoes duplication, one copy generally loses its function and goes extinct, in a process known as nonfunctionalisation (Nei and Roychoudhoury, 1973, Petrov and Hartl, 2000). Alternatively, one of the duplicate paralogues may evolve a new function. This process is called neofunctionalisation (NF). Within the classic model of duplicate gene fates, as proposed by Susumu Ohno in 1970, NF was classified in terms of whether the functional change arises through regulatory or coding functionality (Ohno, 1970). *SULF* falls into both of these categories, having arisen from coding sequence changes relative to its original paralogue, but acting in a regulatory fashion. This has led to the proposal of sRNA-mediated neofunctionalisation (SNF) as a third type of NF, with *SULF* as the exemplar case.

SNF elements may underpin undetected genetic barriers

Previous chapters have identified barriers based on the distribution of SNPs. However, differential function of SNF loci is expected to be based on the presence or absence of the IR, rather than allelic variation. In Chapter 3, the region harbouring the *SULF* IR was only detected because SNPs in adjacent regions showed monophyletic distributions. At 1 kb window resolution, *SULF* was absent, reflecting low read depth (Figure 3.19). If some genetic barriers are underpinned by SNF loci, they may not be detectable through mapping-based approaches. Approaches detailed in this chapter aim to complement the characterisation of barrier loci, by facilitating the detection of SNF candidate IRs.

A bioinformatic approach for detecting SNF candidate loci

To test whether SNF might be involved in broader phenotypic diversity within *Antirrhinum*, a bioinformatic approach, consisting of six filtering steps, was developed in discussion with Simon Moxon and Leighton Folkes. The analyses that follow were all carried out be me. Leighton has separately developed a computational tool for carrying out a similar analysis (Folkes *et al.*, manuscript in preparation). This tool, named SNF, is available at https://github.com/LF-Bioinformatics/SNF.

I will now outline the six filtering steps used to detect SNF candidate IRs. Each step is titled with a criterion – these criteria reflect the *SULF*-like attributes being tested.

1) *SNF candidate loci should be detectable as genomic IRs*

Firstly, if other *SULF*-like SNF loci are present in the genome of *Antirrhinum majus*, then they should be detectable by scanning for IR sequences. I will identify all genomic IRs within eight genome assemblies using Inverted Repeat Finder (IRF) (Warburton *et al.*, 2004). Compared to other IR detection tools, IRF is effective at identifying IRs that contain long spacer regions, like *SULF* (Jia *et al.*, 2022).

2) *SNF candidate IRs should produce 21 nt sRNAs*

To look for functionally relevant IRs, I will next map sRNAs extracted from different *Antirrhinum* species to all IRs from those species' assemblies, using Bowtie (Langmead *et al.*, 2009). IRs will then be filtered based on their sRNA mapping profiles. IRs with low sRNA depth (< 20 mapping sRNAs) will be discarded. To capture *SULF*-like candidates, IRs will be retained if the mode mapped sRNA length is 21 nt.

3) SNF candidate IRs should have target genes

To predict whether IRs have target genes, candidate IR sequences will be compared against predicted coding sequences (CDSs) using local BLASTN searches. In this context, targets refer not to complementary mRNA sites as in miRNA literature (Agarwal *et al.*, 2015), but to CDSs with statistically significant similarity to the longest arm of a given IR. The number of significant BLASTN hits will not be taken into consideration.

4) SNF candidate IRs should show presence / absence relationships between species

SNF candidate IRs are expected to show presence / absence relationships between species. To test this, candidate IRs will be mapped to all available species genome assemblies using Minimap2 (Li, 2021). IRs that are present within all assemblies will be discarded.

IRs meeting criteria 1-4 represent good SNF candidate loci. I also propose an additional two criteria, which may help in characterising how SNF candidate loci function.

5) SNF candidate IRs should show similarity to characterised proteins

Functional information will be inferred by comparing the longest arm of each SNF candidate IR to the NCBI non-redundant protein (nr) database using BLASTX, and examining annotations.

6) SNF candidate IRs may underlie growth habit divergence

If SNF candidate IRs are involved in divergence, then genomic regions harbouring IRs should segregate between divergent *Antirrhinum* species. The characterisation of genomic regions showing growth habit monophyly in Chapter 4 presents an opportunity to test whether genomic regions containing SNF candidate IRs show evidence of segregation between alpine and ruderal species, provided that a given SNF candidate IR is present within the *A. majus* reference genome. By mapping SNF candidate IRs to the *A. majus* reference genome, and comparing their locations against the 427 regions showing growth habit monophyly, I will examine whether any identified IRs might underpin genetic barriers involved in growth habit.

Aim of this work

This work presents a case study for the bioinformatic detection of SNF candidate IRs within genome assemblies of eight *Antirrhinum* species. Using sRNA libraries derived from tissue from *A. m. pseudomajus*, *A. m. striatum*, *A. molle*, and *A. sempervirens*, detection steps 1-5 will be carried out. Each step will sequentially filter genomic IRs, using predefined criteria based on *SULF*. Taken together, these analyses will test the effectiveness of the SNF candidate detection pipeline, and provide preliminary evidence as to whether SNF is a common phenomenon within *Antirrhinum*. If identified SNF candidate IRs are also present within the *A. majus* reference genome, their location will be checked against the list of monophyletic islands identified in Chapter 4, to detect whether IRs may be involved in growth habit divergence.

Results

Results: Less than 0.5 % of the *Antirrhinum* genome contains inverted repeats

To quantify genomic IRs within *Antirrhinum* species, including those which contain long spacer regions between repeats, Inverted Repeat Finder was run on eight genome assemblies. Figure 5.3 shows histograms of IR counts from eight genome assemblies, grouped according to IR length. A total of 257,909 IRs were identified across all assemblies. IRs accounted for between 0.11 % and 0.46 % of total genomic sequence within each assembly. 27.5 – 28.9 % of IRs were within the *SULF* size class, between 1,000 nt and 2,000 nt in length. This suggests that *SULF*-like long IRs represent a minority of total genomic IRs. 51.8 – 56.4 % of IRs were small (< 1000 nt). Each assembly contained at least one very large IR (> 20 kb).

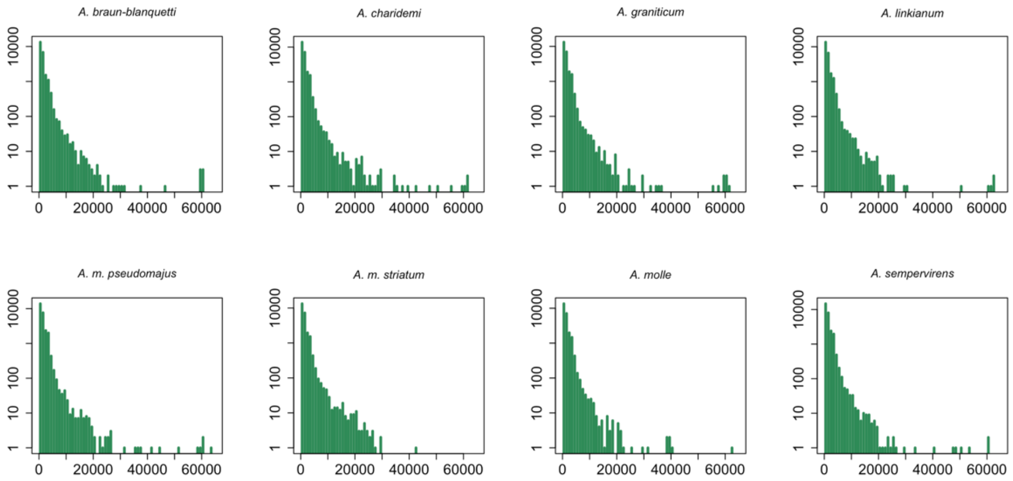


Figure 5.3: Histograms of IR sizes in eight species assemblies

Histograms showing the frequency of different size classes of IRs within the eight species assemblies.

Results: 0.89 % of IRs primarily yield 21 nt sRNAs

To determine the proportion of genomic IRs that yield sRNAs, sRNAs were extracted from petals of three *Antirrhinum* species by Desmond Bradley, and sequenced by Maria-Elena Mannarelli. Two ruderal species, *A. m. pseudomajus* and *A. m. striatum*, and one alpine species, *A. molle*, were included. These three assemblies contained 76,808 IRs. sRNAs from each species were filtered to remove reads < 18 nt and > 28 nt, and were mapped to all IRs from the corresponding species assembly using Bowtie. Figure 5.4 shows the numbers of IRs meeting different filtering criteria. 75,786 IRs had at least one mapping sRNA. A minimum depth threshold of 20 mapped sRNAs was applied, leaving 67,427 mapped IRs in total. 98 % (66,119 / 67,427) of remaining IRs showed a mode sRNA length of 24 nt. 687 IRs had a mode sRNA length of 21 nt. This included the *SULF* hairpin in *A. m. pseudomajus* (Figure 5.5). These "mode 21 nt" IRs represented the second largest size class amongst mapped IRs (Figure 5.6).

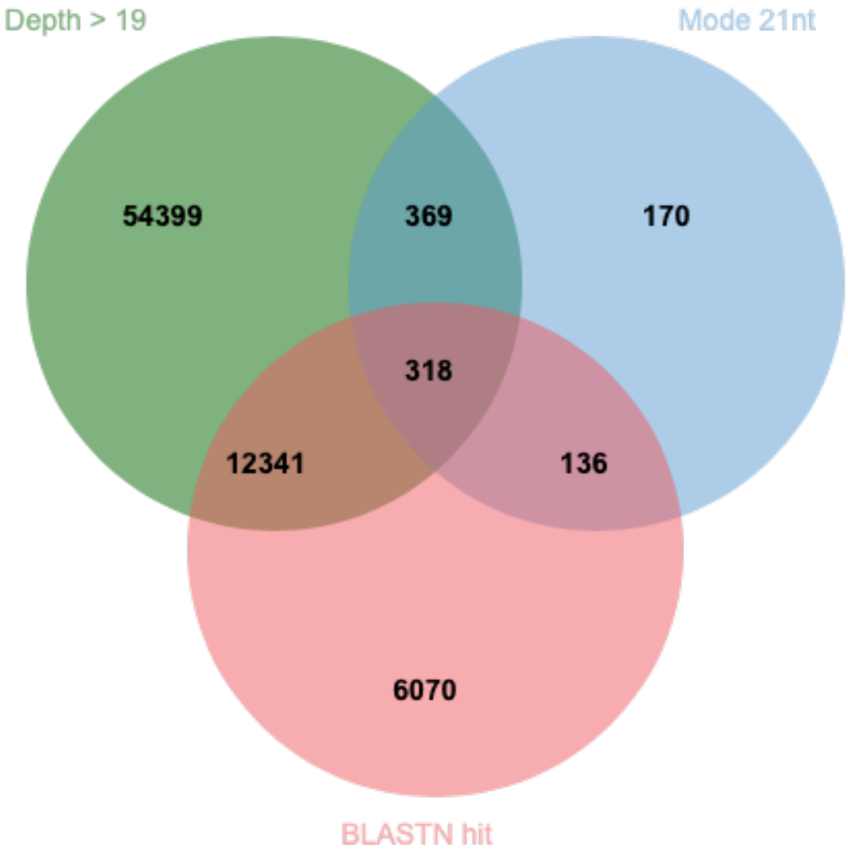


Figure 5.4: Venn diagram showing the characteristics of sRNA-mapped IRs from A. m. pseudomajus, A. m. striatum, and A. molle

"Depth > 19" IRs comprise all IRs with > 19 mapped reads. "Mode 21 nt" IRs comprise all IRs with a mode mapped read length of 21 nt. "BLASTN hit" IRs are all IRs with at least one significant BLASTN hit when compared to predicted CDSs.

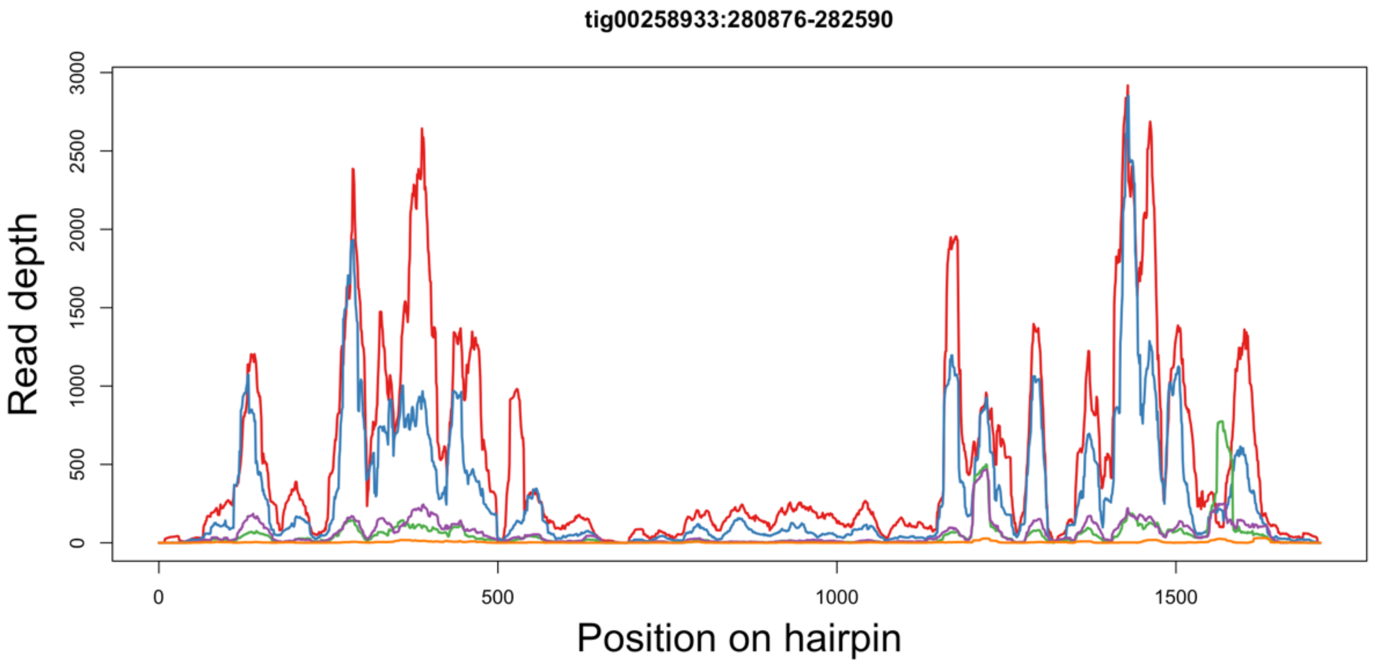


Figure 5.5: sRNA coverage plot of the SULF locus

Read depth of 21-25 nt sRNAs plotted across the *SULF* IR. Each coloured line represents a different sRNA size class. 21 nt sRNA depth is shown in red. The genomic position of *SULF* within the *A. m. pseudomajus* assembly is displayed above the plot, in the format *scaffold:start-end*.

For IRs to be SNF candidates, a potential target gene must be observed. To identify possible targets based on sequence similarity, local BLASTN searches were carried out. The longest arm of each of the 687 mode 21 IRs was compared to predicted CDSs from the *A. m. pseudomajus*, *A. m. striatum*, and *A. molle* assemblies using default BLASTN parameters. This revealed that 318 / 687 mode 21 IRs showed similarity to predicted CDSs, indicating possible target genes. These 318 mode 21 IRs represented the best SNF candidates, and were retained for further analysis.

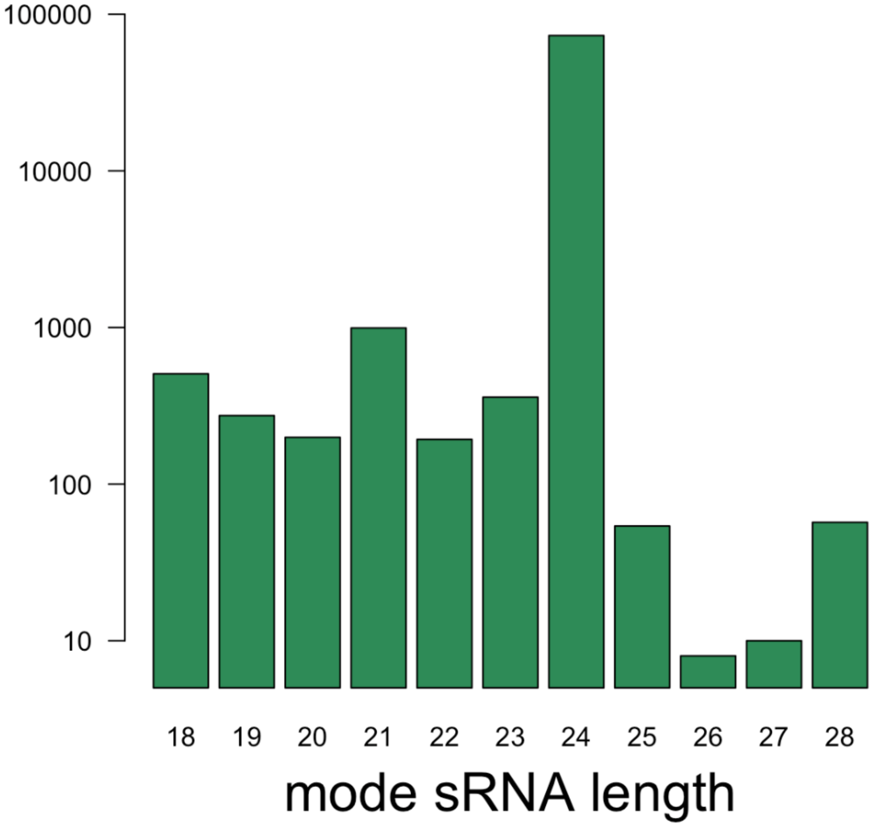


Figure 5.6: Histogram of mode mapped sRNA length

Frequencies of different mode mapped sRNA lengths across the 75,786 IRs from *A. m. pseudomajus*, *A. m. striatum*, and *A. molle*. The 22 IRs with no mapping sRNAs are excluded.

Results: 152 candidate IRs segregate between species

To determine whether SNF candidate IRs are shared between test species, each was mapped to all eight genome assemblies using Minimap2. Two very short (< 200 nt) IRs failed to map to any assemblies. These were removed from the candidate set. Presence / absence patterns of the remaining 316 SNF candidate IRs are summarised in Figure 5.7. 52 % of SNF candidate IRs mapped to all assemblies, suggesting that they are unlikely to be involved in species divergence. 152 IRs were absent within at least one assembly. SNF candidate IRs that show presence / absence relationships between alpine and ruderal species are of particular interest, as they may be involved in growth habit divergence. 45 IRs were absent within both alpine species, *A. molle* and *A. sempervirens*. Reciprocally, 4 IRs were unique to *A. molle*, and 2 were shared between *A. molle* and *A. sempervirens* only.

Results: 40.8 % of IRs have similarity to protein coding genes

The 51 growth-habit-candidate IRs, along with the other 101 IRs segregating between species, were compared to the NCBI nr database using BLASTX. Again, BLAST searches were carried out using the longest arm of each SNF candidate IR. Using default parameters, 62 / 152 SNF candidate IRs had significant similarity to translated proteins (Table 5.1). 34 % of these 62 SNF candidate IRs showed presence / absence according to growth habit.

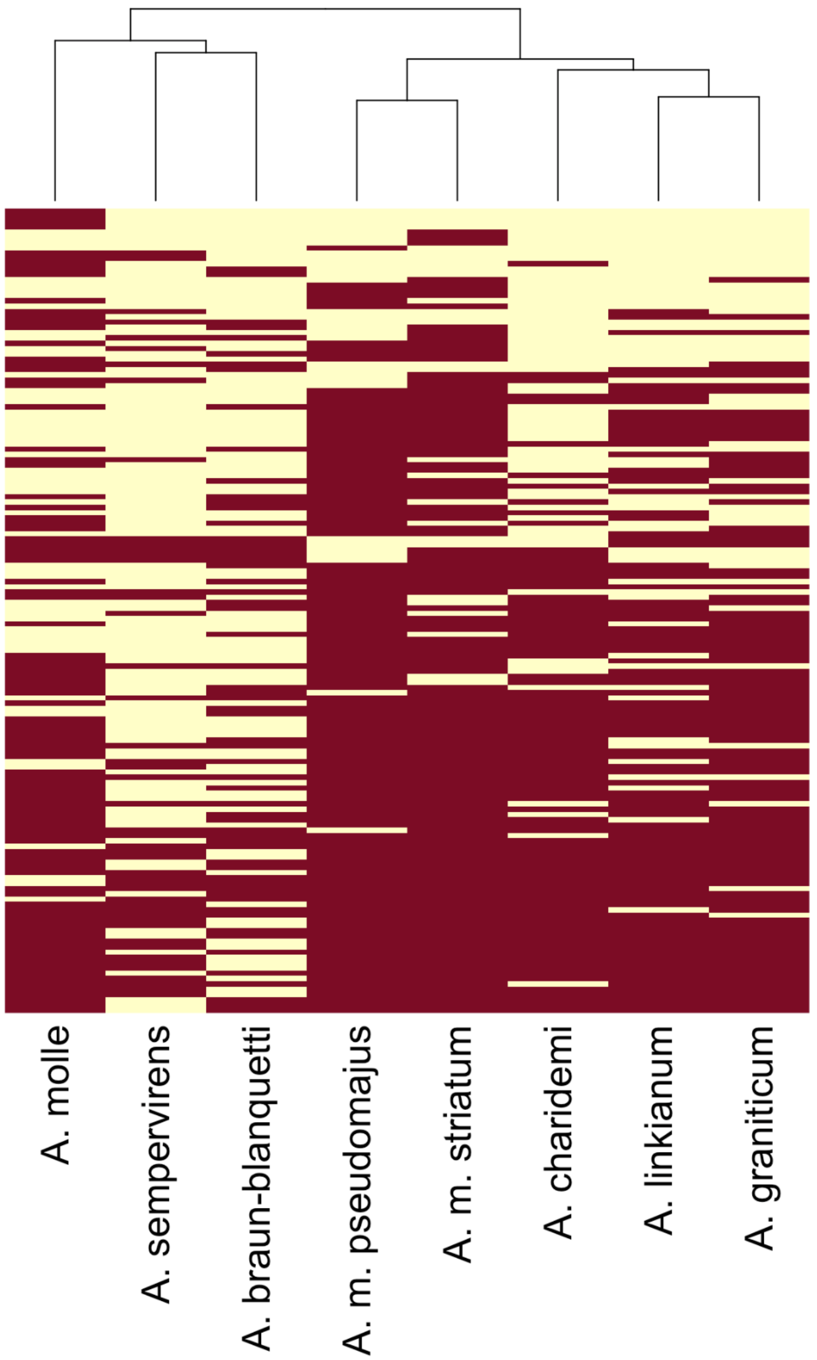


Figure 5.7: Presence / absence heatmap of SNF candidate IRs in species

Two colour heatmap showing presence / absence of 411 segregating SNF candidate IRs (y-axis) identified from *A. m. pseudomajus*, *A. m. striatum*, and *A. molle*. IRs that are present within a given species are coloured red. Results have been clustered based on the number of shared IRs. This is summarised in the UPGMA tree atop the heatmap.

Description	Species	Query coverage	E value	% identity
E3 ubiquitin-protein ligase RING1-like	*Sesamum indicum*	57 %	$2.00e^{-34}$	45.51
sugar porter family MFS transporter	*Serratia marcescens*	81 %	0.001	44.44
lysine-specific demethylase jmj25	*Phtheirospermum japonicum*	99 %	$4.00e^{-24}$	52.1
protein embryonic flower 1	*Phtheirospermum japonicum*	49 %	$1.00e^{-4}$	62.16
protein embryonic flower 1	*Phtheirospermum japonicum*	75 %	$3.00e^{-8}$	56.25
auxin-responsive protein SAUR19-like	*Sesamum indicum*	94 %	$6.00e^{-49}$	83.7
auxin-responsive protein SAUR19-like	*Sesamum indicum*	92 %	$2.00e^{-50}$	81.52
Dynamin-related protein like	*Actinidia chinensis*	76 %	$6.00e^{-34}$	36.97
putative mannitol dehydrognase-like	*Trifolium medium*	84 %	$2.00e^{-28}$	73.85
WAT1-related protein	*Striga hermonthica*	29 %	$5.00e^{-8}$	59.52
F-box/kelch-repeat protein at3g23880	*Phtheirospermum japonicum*	83 %	$6.00e^{-6}$	43.1
pectinesterase 2-like	*Coffea arabica*	84 %	$3.00e^{-11}$	41.3
E3 ubiquitin-protein ligase RING1-like	*Sesamum indicum*	24 %	$9.00e^{-38}$	56.59
Regulator of rDNA transcription protein 15	*Capsicum baccatum*	97 %	$5.00e^{-14}$	96.67
WAT1-related protein	*Striga hermonthica*	29 %	$5.00e^{-8}$	59.52
F-box/LRR-repeat protein at4g14096	*Phtheirospermum japonicum*	59 %	$2.00e^{-7}$	40.48
putative mannitol dehydrognase-like	*Trifolium medium*	84 %	$2.00e^{-28}$	73.85
F-box/kelch-repeat protein at3g23880	*Phtheirospermum japonicum*	83 %	0.001	37.93
pectinesterase 2-like	*Coffea arabica*	85 %	$2.00e^{-10}$	42.39
auxin-induced protein 15A	*Sesamum indicum*	98 %	$6.00e^{-43}$	81.52
auxin-induced protein 15A	*Sesamum indicum*	98 %	$7.00e^{-45}$	84.15
auxin-induced protein 15A	*Sesamum indicum*	87 %	$1.00e^{-46}$	77.17
protein DELAY OF GERMNATION 1-like	*Sesamum indicum*	78 %	$7.00e^{-41}$	54.55
F-box/kelch-repeat protein at3g23880	*Phtheirospermum japonicum*	85 %	$4.00e^{-4}$	43.1
putative B3 domain-containing protein At5g66980	*Sesamum indicum*	46 %	$1.00e^{-9}$	58

Table 5.1 (previous page): BLASTX hits for 25 SNF candidate IRs.

BLASTX hits against the nr database for 25 SNF candidate IRs. The longest repeat arms from 152 SNF candidate IRs were compared to the NCBI nr database using BLASTX. 62 / 152 IRs had significant BLASTX hits. 37 annotations were hypothetical / uncharacterised proteins. These have been omitted, leaving 25 with similarity to known proteins. Query coverage refers to the percentage of the repeat arm with sequence similarity to the protein hit. E-value reports the statistical significance of the BLAST hit. % identify shows the percentage of matching amino acids between the translated repeat arm, and the protein hit.

Results: Two SNF candidate IRs are localised within monophyletic islands

To test whether SNF candidate IRs resided in genomic regions that have been observed to segregate between alpine and ruderal *Antirrhinum* species, IRs were mapped to the *A. majus* reference genome using Minimap2. 106 / 152 candidates mapped to the reference genome. To search for monophyletic signatures, the eight species dataset analysed in Chapter 4 was utilised. 1 kb window trees, with 900 bp overlaps, were constructed across the two 50 kb windows harbouring the SNF candidate regions showing growth habit segregation. Two regions yielded trees that were doubly monophyletic for growth habit. The first region showed very low sRNA depth, suggesting that the SNF is unlikely to be expressed in the petal tissue analysed (Figure 5.8). The second region, on chromosome 7, showed a monophyletic island size of around 10 kb. The IR, spanning 1,788 nt and identified within the *A. molle* assembly, showed 21 nt sRNA peaks of moderate depth on either arm (Figure 5.9). A BLASTX search comparing the longest arm of this IR to the nr database suggested that it contains an F-box domain. This is not shown in Table 5.1, as the most significant hit corresponded to a hypothetical protein.

In total, this analysis detected 152 SNF candidate IRs passing criteria 1-4. Of these, 25 had similarity to characterised proteins within the NCBI nr database. Two SNF candidate IRs were seen to be localised within genomic regions that are divergent between alpine and ruderal *Antirrhinum* species, although one showed very low sRNA depth, and may be an artifact.

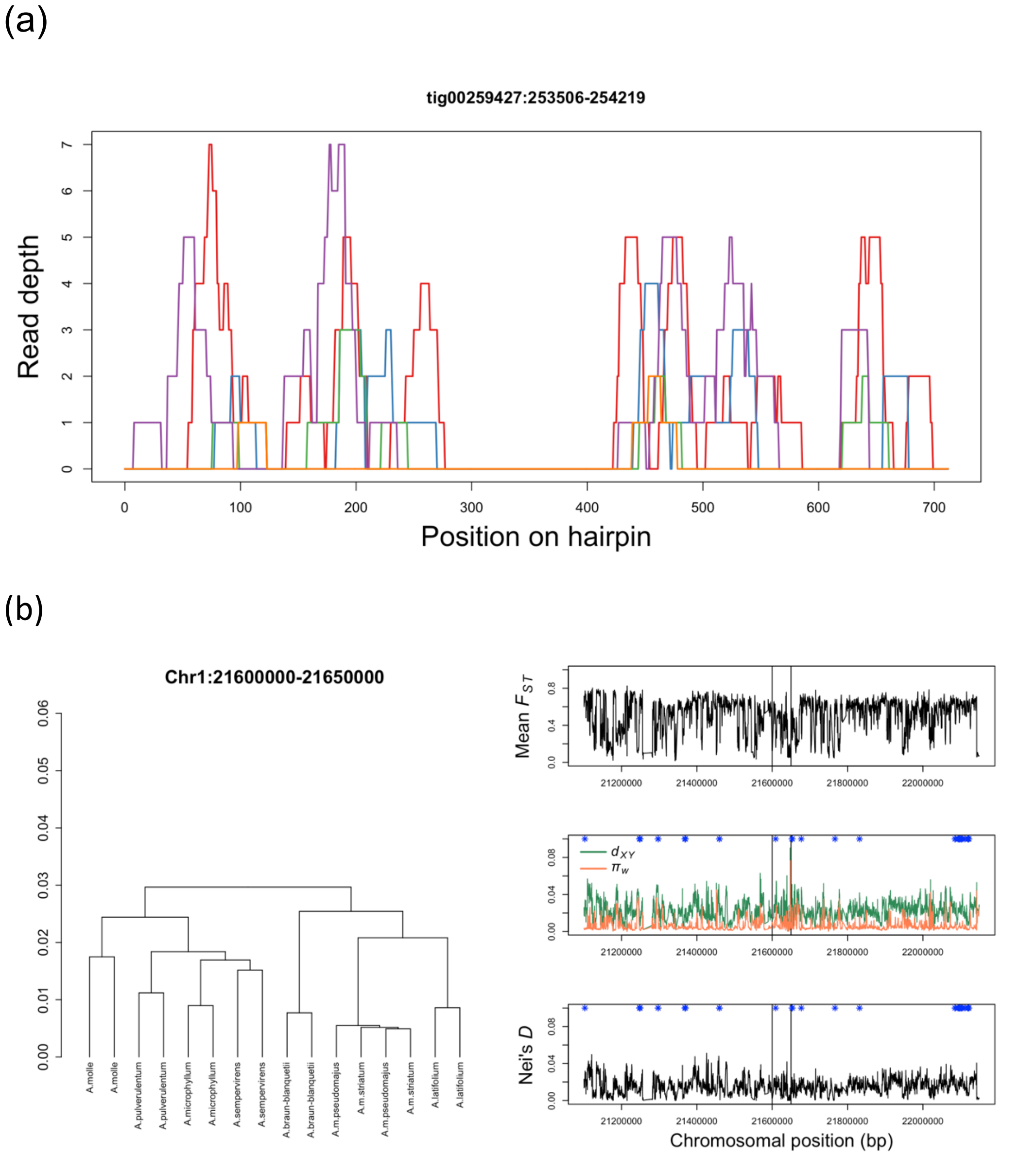

(a)
tig00259427:253506-254219
Read depth
0
1
2
3
4
5
6
7
Position on hairpin
0
100
200
300
400
500
600
700
(b)
Chr1:21600000-21650000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.pulverulentum
A.pulverulentum
A.microphyllum
A.microphyllum
A.sempervirens
A.sempervirens
A.braun-blanquetii
A.braun-blanquetii
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
A.latifolium
A.latifolium
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
0.00
0.04
0.08
21200000
21400000
21600000
21800000
22000000
Chromosomal position (bp)

Figure 5.8 (previous page): sRNA coverage plot and monophyletic island plots for a candidate monophyletic SNF candidate IR.

(a) Read depth of 21-25 nt sRNAs plotted across the *candidate* IR. Each coloured line represents a different sRNA size class. 21 nt sRNA depth is shown in red. The genomic position of *the candidate IR* within the A. m. pseudomajus assembly is displayed above the plot, in the format scaffold:start-end. (b) Plots showing diversity statistics across a ca. 1 Mb region containing the candidate IR. (left) Summarising 50 kb window tree for the whole monophyletic region. (right, top) Mean F_{ST} across the monophyletic region from comparisons of all populations, calculated in 1 kb windows with 900 bp overlaps. (right, middle) Mean D_{XY} (green) from all population comparisons, and mean πw (orange) from all populations, summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. (right, bottom) Nei's D (D_{XY} - π_W) summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. Blue asterisks indicate the locations of 1 kb window trees that are monophyletic for alpine and ruderal growth habits. The black rectangle represents the limits of the originating 50 kb monophyletic region.

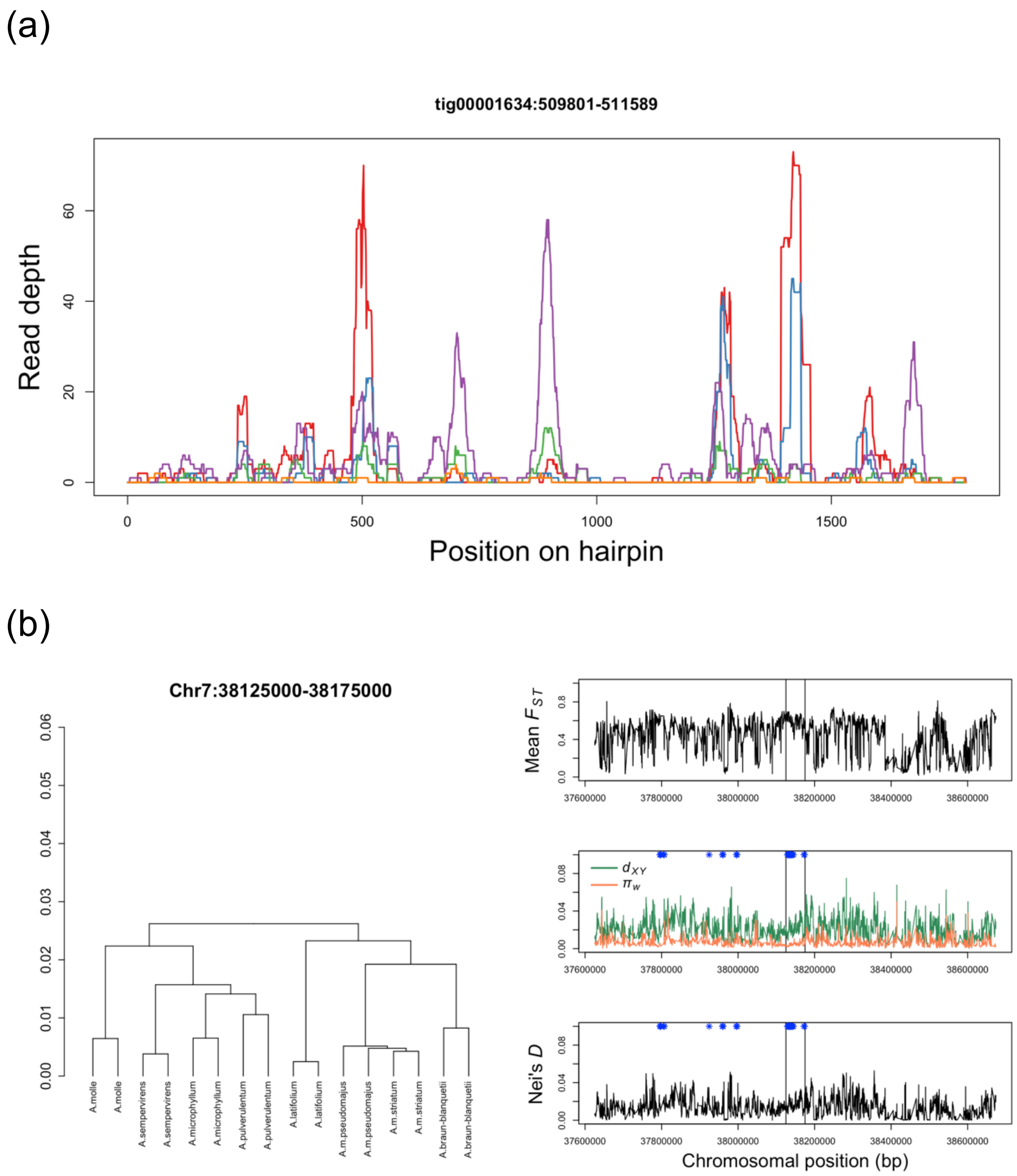

(a)
tig00001634:509801-511589
Read depth
0
20
40
60
0
500
1000
1500
Position on hairpin
(b)
Chr7:38125000-38175000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.sempervirens
A.sempervirens
A.microphyllum
A.microphyllum
A.pulverulentum
A.pulverulentum
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
A.m.striatum
A.braun-blanquetii
A.braun-blanquetii
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
0.00
0.04
0.08
37600000
37800000
38000000
38200000
38400000
38600000
Chromosomal position (bp)

Figure 5.9 (previous page): sRNA coverage plot and monophyletic island plots for a candidate monophyletic SNF candidate IR.

(a) Read depth of 21-25 nt sRNAs plotted across the *candidate* IR. Each coloured line represents a different sRNA size class. 21 nt sRNA depth is shown in red. The genomic position of *the candidate IR* within the A. m. pseudomajus assembly is displayed above the plot, in the format scaffold:start-end. *(b)* Plots showing diversity statistics across a ca. 1 Mb region containing the candidate IR. (left) Summarising 50 kb window tree for the whole monophyletic region. (right, top) Mean F_{ST} across the monophyletic region from comparisons of all populations, calculated in 1 kb windows with 900 bp overlaps. (right, middle) Mean D_{XY} (green) from all population comparisons, and mean π_W (orange) from all populations, summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. (right, bottom) Nei's D (D_{XY} - π_W) summarised in 1 kb windows with 900 bp overlaps across the monophyletic region. Blue asterisks indicate the locations of 1 kb window trees that are monophyletic for alpine and ruderal growth habits. The black rectangle represents the limits of the originating 50 kb monophyletic region.

Results: Analysis of sRNA from *A. m. pseudomajus* and *A. sempervirens* reveals more SNF candidate IRs

To expand the SNF search, the sRNA extraction and sequencing was carried out in triplicate on samples from *A. m. pseudomajus* and *A. sempervirens*. sRNAs were extracted from four tissue types; leaves, flower buds, petals, and shoots. To capture as many candidates as possible, replicate datasets were considered individually in the first instance. A total of 54,947 IRs were identified within the *A. m. pseudomajus* and *A. sempervirens* genome assemblies. Mapping the new sRNAs, 44,397 IRs had at least 20 mapping sRNAs in one or more replicates from one or more tissues. 11,150 IRs had predominately 21 nt sRNAs mapped in one or more replicates. To look for possible target genes, the longest arm of each of these IRs was compared to predicted *A. m. pseudomajus* or *A. sempervirens* CDSs using local BLASTN. 982 / 11,150 IRs had possible CDS targets based on sequence similarity (Figure 5.10). Again, mode 21 IRs represented the second largest size class of mapped IRs (Figure 5.11).

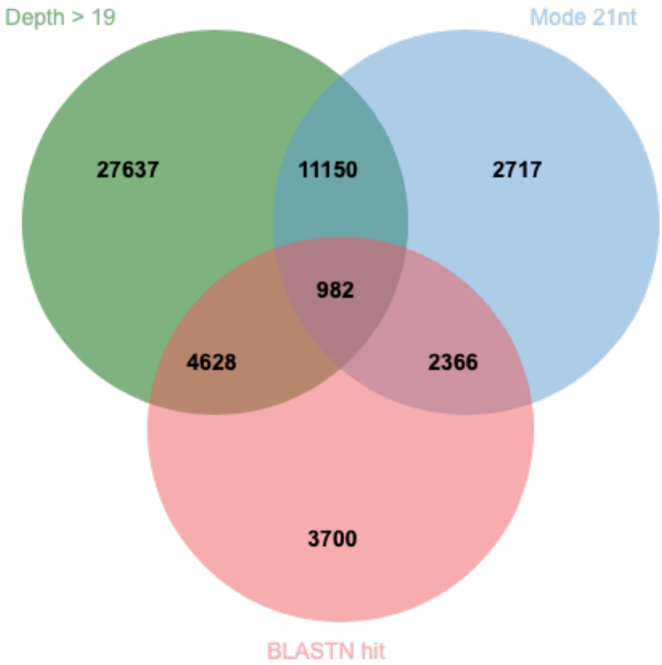


Figure 5.10: Venn diagram showing the characteristics of sRNA-mapped IRs from A. m. pseudomajus, and A. sempervirens

"Depth > 19" IRs comprise all IRs with > 19 mapped reads. "Mode 21 nt" IRs comprise all IRs with a mode mapped read length of 21 nt. "BLASTN hit" IRs are all IRs with at least one significant BLASTN hit when compared to predicted CDSs.

Of the 980 IRs that mapped to at least one assembly, 487 were present in all assemblies. 493 IRs showed evidence of presence / absence in one or more assemblies. 493 SNF candidate IRs was too large a number to easily carry out BLASTX searches. To narrow this set down, only those candidates which showed mode 21 nt sRNAs within all tissues were retained for BLASTX comparisons. A total of 35 SNF candidate IRs showed a mode mapped sRNA length of 21 in all tissues. 27 of these gave hits when compared against the NCBI nr database using BLASTX (Table 5.2).

This analysis detected 493 SNF candidate IRs passing criteria 1-4. Of the 35 of these that were selected for BLASTX searches, nine showed similarity to characterised protein domains (criterion 5). Comparison to known monophyletic regions was not carried out for this set.

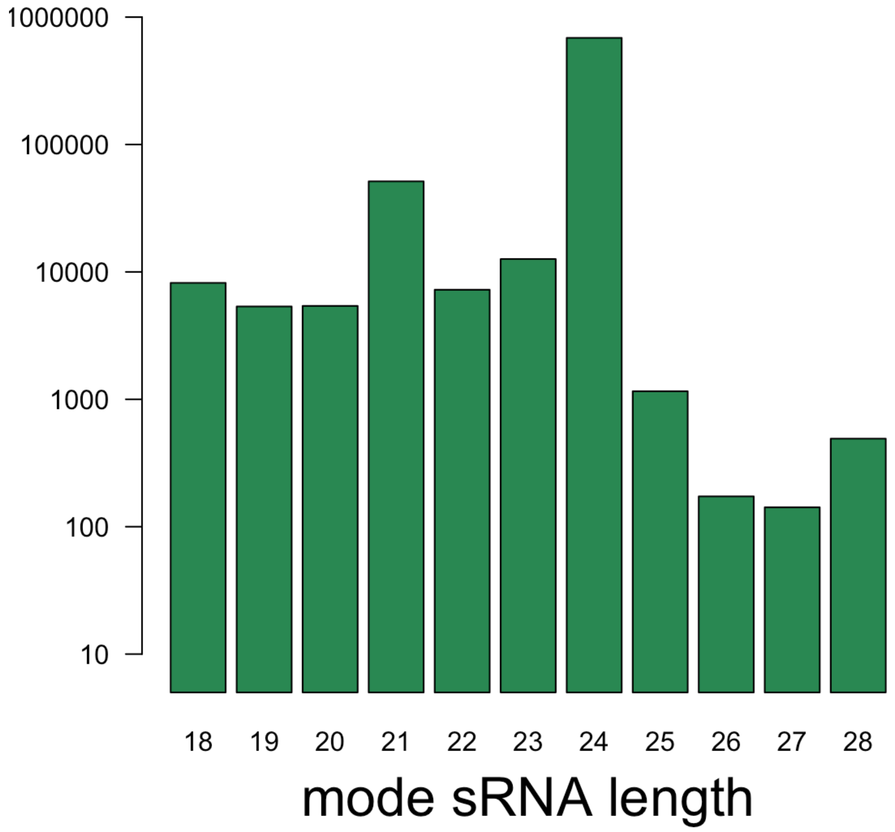


Figure 5.11 Histogram of mode mapped sRNA length

Frequencies of different mode mapped sRNA lengths across the 54,947 IRs from *A. m. pseudomajus* and *A.* sempervirens. The 359 IRs with no mapping sRNAs are excluded.

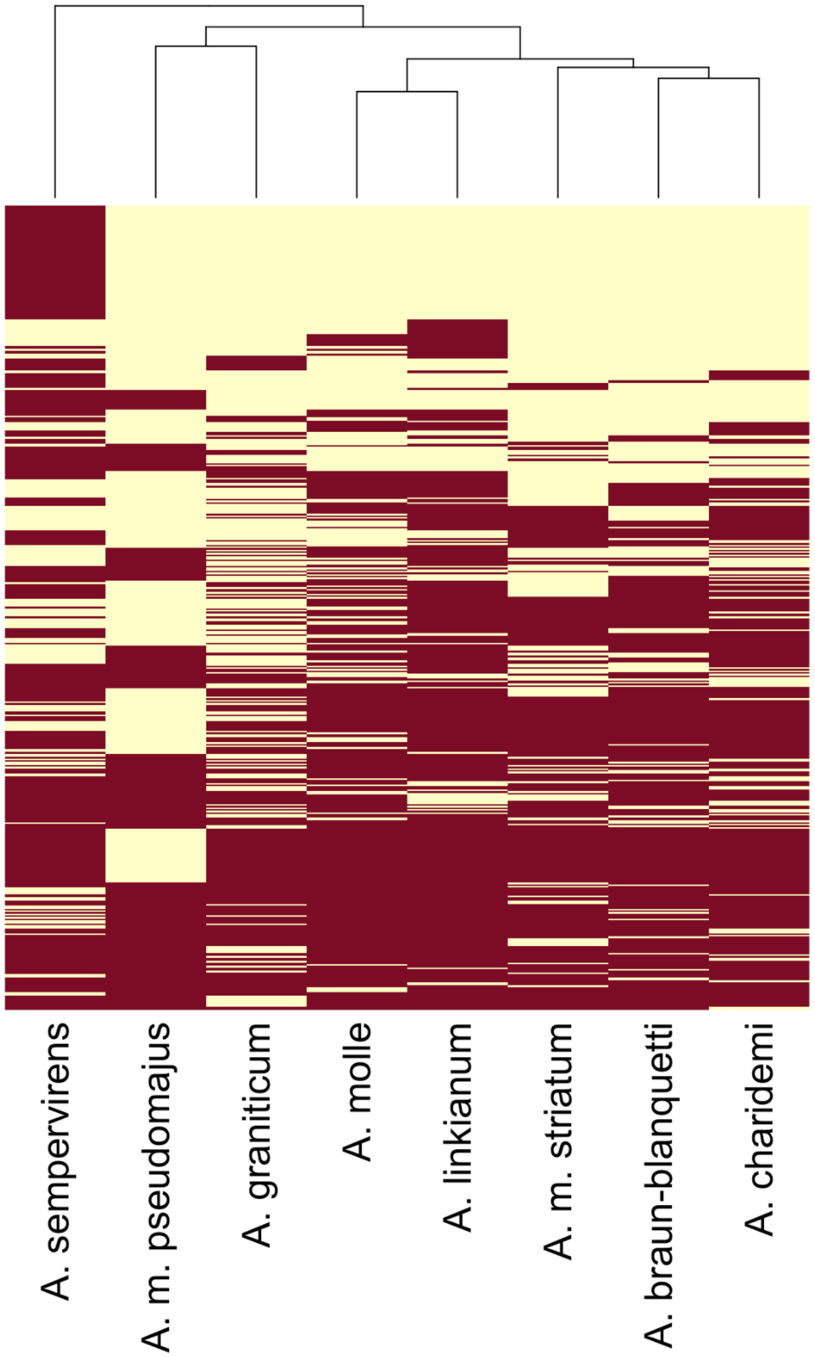


Figure 5.12: Presence / absence heatmap of SNF candidate IRs in species

Two colour heatmap showing presence / absence of 411 segregating SNF candidate IRs (y-axis) identified from *A. m. pseudomajus* and *A. sempervirens*. IRs that are present within a given species are coloured red. Results have been clustered based on the number of shared IRs. This is summarised in the UPGMA tree atop the heatmap.

Description	Species	Query coverage	E value	% identity
F-box/kelch-repeat protein at3g23880	*Phtheirospermum japonicum*	83 %	6.00e^{-6}	43.1
putative disease resistance protein RGA1	*Sesamum indicum*	25 %	3.00e^{-10}	59.68
mediator of RNA polymerase II transcription subunit 25 isoform X1	*Sesamum indicum*	50 %	3.00e^{-9}	61.97
putative F-box/LRR-repeat protein At5g41840	*Sesamum indicum*	96 %	4.00e^{-29}	52.48
F-box/kelch-repeat protein at3g23880	*Phtheirospermum japonicum*	83 %	6.00e^{-6}	43.1
pectinesterase-like	*Coffea arabica*	38 %	4.00e^{-6}	57.14
E3 ubiquitin-protein ligase RING1-like	*Sesamum indicum*	57 %	2.00e^{-34}	45.51
putative F-box/LRR-repeat protein At5g41840	*Sesamum indicum*	99 %	1.00e^{-32}	56.74
Dynamin-related protein like	*Actinidia chinensis*	76 %	6.00e^{-34}	36.97

Table 5.2: BLASTX hits for 9 SNF candidate IRs.

BLASTX hits against the nr database for nine SNF candidate IRs. The longest repeat arms from 35 SNF candidate IRs were compared to the NCBI nr database using BLASTX. 26 / 35 IRs had significant BLASTX hits. 17 annotations were hypothetical / uncharacterised proteins. These have been omitted, leaving nine with similarity to known proteins. Query coverage refers to the percentage of the repeat arm with sequence similarity to the protein hit. E-value reports the statistical significance of the BLAST hit. % identify shows the percentage of matching amino acids between the translated repeat arm, and the protein hit.

Discussion

0.5 % of surveyed IRs met all SNF candidate criteria

The primary aim of this chapter was to develop and apply a bioinformatic pipeline, to serve as a case study for the detection of SNF candidate IRs. To do this, I defined four criteria that a *SULF*-like SNF candidate IR should meet. These analyses of *A. m. pseudomajus*, *A. m. striatum*, *A. molle*, and *A. sempervirens* have identified 645 SNF candidate IRs, out of the 131,755 IRs present in the four species genome assemblies. 152 SNF candidate IRs, including *SULF*, were between 1,000-2,000 nt long (Figure 5.13). 206 SNF candidate IRs were larger than 2,000 nt, with the largest IR being 16,552 nt. The theoretical maximum size of a SNF locus depends on the biological processes that can give rise to IRs. Single gene duplication can arise through transposon activity (Wang, Y., Wang, X, and Paterson, 2012). Transposons spanning over 20 kb have been characterised in a range of eukaryotic systems (Arkhipova and Yushenova, 2019). This raises the possibility that an IR generated through multiple transposon mediated gene duplications might be much larger than 2 kb. Presumably, a very large precursor sRNA hairpin would be less efficiently processed by DICER-LIKE, but this is unclear. 38.5 % of SNF candidate IRs were < 1000 nt long. These likely include miRNA precursors (Thakur *et al.*, 2011). If SNF IRs represent "immature" miRNA loci, then some of these small IRs may be older SNF loci, which have lost some of their redundant sequence similarity to their paralogue. Characterised miRNA loci could be excluded by comparing small IRs to the miRBase (Griffiths-Jones *et al.*, 2006) miRNA precursor database, but this is likely to represent only a fraction of all miRNAs in *Antirrhinum*.

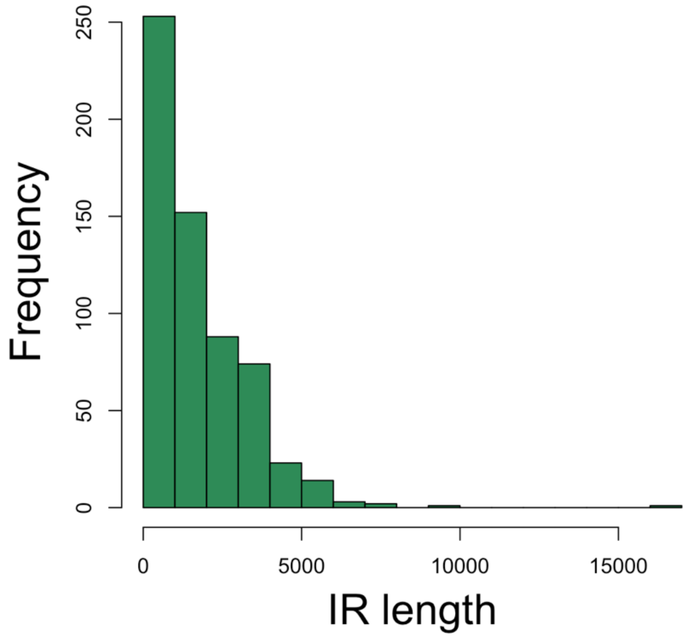


Figure 5.13: Histogram of lengths of all SNF candidate IRs.

Frequency of different IR lengths within the SNF candidate IRs. The large outlying IR was 16,552 nt.

SNF candidates may target F-box proteins, and auxin responsive proteins

A total of 152 SNF candidate IRs were compared to the NCBI nr database using BLASTX. Of these, 34 showed significant BLASTX hits to characterised protein domains – 25 from the first analysis, and nine from the second. Eight of the 34 were predicted to contain F-box domains. Five were of the kelch-box type, and three of the LRR-repeat type. F-box proteins are diverse, and implicated in a range of developmental processes including floral organ development, hormone signal transduction, photoperiodism, stress responses, and metabolism (Zhang *et al.*, 2019). The functional diversity of F-box proteins makes it challenging to speculate on their significance as SNF targets. However, possible roles in development are compelling, particularly as one SNF candidate IR containing an F-box domain was shown to reside in a genomic island that is monophyletic for growth habit. However, the fact that this IR was detected in *A. molle* (an alpine species), but also present within the reference genome (a ruderal inbred line), suggests that it may not be a true SNF locus, which should show presence / absence.

The second most common similarity was to auxin responsive proteins, including the auxin responsive protein SAUR19-like. SAUR (small auxin upregulated RNA) factors are a family of proteins characterised by fast turnaround in response to auxin signalling (Stortenbeker and Bemer, 2019). *SAUR* genes can be rapidly induced in response to a range of stimuli, but their transcripts and proteins have very short half-lives (McClure and Guilfoyle, 1989; Newman *et al.*, 1993; Knauss *et al.*, 2003). To date, no studies have directly implicated sRNAs in *SAUR* transcript degradation, although auxin response factors involved in modulating auxin induced genes are known to be regulated by miRNAs (Mallory, Bartel, D., Bartel, B., 2005).

Evidence presented here suggests that SNF candidate IRs identified in different species show similar functional annotations. However, interpretation of this is limited because not all SNF candidate IRs have been characterised. A greater number of annotations could be obtained by running BLASTX searches for all 493 segregating IRs in the *A. m. pseudomajus* / *A. sempervirens* dataset, rather than just the 35 which showed mode 21 nt sRNAs in all tissues. However, this proved challenging to automate. A preliminary attempt to derive functional predictions using eggNOG-mapper in genomic mode returned hits for 104 / 493 IR arms. This likely represents a better option than using BLASTX, being significantly simpler to run for hundreds of input sequences.

Improving the SNF candidate detection pipeline

Work presented here is intended as a case study, for future development of *in silico* approaches to characterise SNF candidate loci. It is important, therefore, to consider the effectiveness of the approach. A challenge of detecting SNF candidates is that currently only one locus, *SULF*, serves as an exemplar or signature. Thus, the properties of SNF candidates were poorly defined, and it was important that detection criteria were reasonably relaxed. To capture as many SNF candidate IRs as possible, it is important that genomic IRs be detected effectively. An estimated

52.6 % of the most recently published *Antirrhinum majus* genome is repetitive (Li *et al.*, 2019). Running IRF on the reference genome revealed that 4.6 % of genomic sequence corresponds to IR sequences. The percentage of repetitive sequence within each of the eight species assemblies was found to be between 45.2 % and 49.4 %. This is slightly less than the reference genome. However, the reference genome is a published, chromosome level assembly, and is therefore likely to be more complete. Reported genomic IR content within the species assemblies was less than 0.5 %. This tenfold reduction suggests that the number of IRs detected may vary substantially depending on the quality and completeness of the assembly used.

A higher depth filter earlier on in the pipeline would reduce noise. Over 98 % of genomic IRs showed predominantly 24 nt mapping sRNAs, and were therefore disregarded. However, many of these IRs may have also had 21 nt sRNAs mapped. A more lenient search could include IRs where the depth of 21 nt sRNAs is similar to the depth of 24 nt sRNAs. The *SULF* IR showed a characteristic pattern of sRNA mapping, with a tall, broad 21 nt sRNA peaks on each arm. This distribution fits the biological expectation that processed sRNAs should originate from sequence showing complementarity to their targets. The approach detailed here lacks any means of discriminating based on read distribution. Ideally, the distribution of 21 nt sRNAs across each IR would be numerically defined, and IRs with sporadically mapping sRNAs excluded. A simple way to achieve this would be to filter based on depth per position, or by applying a threshold sRNA coverage. Both approaches would exclude IRs with sporadically mapping sRNAs. The program ShortStack (Axtell, 2013) can be used to improve sRNA mapping data by carrying out *de novo* detection of sRNA clusters, and quantifying the RNA folding dynamics of the underlying IR sequence, to infer the likelihood that an expressed IR would spontaneously fold *in vivo*. Shortstack is incompatible with the SNF detection pipeline detailed here, as it requires that sRNA libraries are mapped to a reference genome, rather than mapping to extracted IR sequences. Future analyses could move towards whole genomic sRNA mapping, with minimal changes to the downstream analyses. Local BLASTN searches of IR arms against predicted CDSs were effective at filtering many IRs that were unlikely to be SNF candidates. However, whether 21 nt sRNAs mapped directly to the predicted target was not investigated. A more sophisticated approach to target prediction would map sRNAs from each hairpin to its predicted target CDS, and filter based on depth / mode sRNA length. A similar approach is utilised by TarHunter to detect miRNA target sequences (Ma *et al.*, 2018).

Software for the detection of SNF candidate IRs

Analyses presented here involve the use of a range of software tools, and custom processing scripts, which require experience in bioinformatics to use effectively. If SNF is to be studied more widely, it is important that tools are developed which can be used by non-specialists. Leighton Folkes has developed a command line tool integrating the different steps of this identification process, which we developed together in discussion (Folkes *et al.*, manuscript in preparation). The SNF implementation is broadly similar to the approach outlined here, but can be deployed on compatible machines with minimal effort. Because SNF analyses one assembly at a time, additional *ad hoc* analyses are required to test whether detected IRs show presence / absence relationships between assemblies.

Recently, Jia *et al.* (2022) established LIRBase, an online database of long inverted repeats from eukaryotic genomes. This approach shows strong parallels to the analyses presented here, and with the SNF tool. Jia *et al.* used Inverted Repeat Finder to predict inverted repeats in 424 eukaryotic genome assemblies. They specifically identified long inverted repeats by filtering all IRs where both arms were shorter than 400 nt. To detect hairpin sRNAs, the user has the option of providing sRNA data, which is aligned to detected IRs using Bowtie. Possible downstream analyses include differential expression analysis of sRNAs using DESeq2, prediction of target genes by mapping IRs to CDSs using Bowtie, and visualisation of IR structure using RNAfold (Lorenz *et al.*, 2011). IRs can be filtered on the depth of mapped sRNAs, but also on their position, circumventing the issues of sporadic read mapping addressed earlier. The LIRBase suite is a promising resource for future analyses of SNF, which should be explored further.

6: Discussion

Summary of the work presented in this thesis

This work set out to characterise the genetic barriers involved in the divergence of *Antirrhinum* species. Specifically, two divergence events were considered; first between closely related *A. m. pseudomajus* and *A. m. striatum* populations in a shared environment, and the second, more ancient, divergence of *Antirrhinum* species showing distinct growth habits. Prior to analyses, two genetic barriers were hypothesised. Within a shared environment, I hypothesised that intrinsic epistatic barriers might be in place; the reproductive barrier exists due to genetic incompatibilities arising from epistasis. Across distinct environments, with different ecological conditions, I hypothesised the existence of differentially adaptive barriers, underpinned by genes that confer an adaptive advantage within a specific environment.

Work presented here has inferred the genetic basis of barriers in both experimental systems using the grouping-tree-scan approach to identify monophyletic islands. By leveraging pools of genomic DNA from multiple populations, I have identified panels of genomic regions showing signatures of consistent between-population segregation with the phenotypes under investigation. Additionally, in response to the observation that genetic barriers might be underpinned by sRNA loci showing presence / absence distributions between populations, I have presented a case study for identifying similar loci.

In the case of *A. m. pseudomajus* and *A. m. striatum*, I analysed the six identified monophyletic islands in light of SNP and RNAseq data, and concluded that all are likely to have a role in controlling flower colour. This implies, though does not directly demonstrate, that the genetic barrier between *A. m. pseudomajus* and *A. m. striatum* is underpinned by incompatibilities between colour genes.

Analysing eight species showing distinct growth habits, I demonstrated that monophyletic signatures are restricted to a set of regions encompassing 3.1 % of the genome, none of which were detectable using conventional genome scans. Estimating the sizes of the monophyletic islands within each monophyletic region, I concluded that they are consistent with barriers to gene flow or allele sharing, but are generally significantly larger than would be expected by the prevailing parallel allele recruitment hypothesis.

Carrying out a search for *SULF*-like inverted repeat loci, which may have evolved through SNF, revealed a total of 645 IRs with properties similar to *SULF*. These represent the best SNF candidate loci, but further characterisation is required to test whether they are involved in phenotypic divergence.

The reproductive barrier between *A. m. pseudomajus* and *A. m. striatum* likely reflects coadaptation of colour loci

In the Introduction I postulated that an intrinsic epistatic barrier should have three properties: equivalent solutions to a shared problem, distinct solutions that operate irrespective of environment, and maladaptive consequences for mixed strategies. Characterising the divergent genomic regions between *A. m. pseudomajus* and *A. m. striatum* suggests that all three of these criteria have been met. The shared problem faced by ruderal *Antirrhinum* species is one of pollinator attraction in competitive environments. Colour patterns in *A. m. pseudomajus* and *A. m. striatum* are solutions to the pollinator attraction problem, showing adaptation that is likely to make them effective floral guides to their pollinating bumblebees. The two subspecies show distinct but equivalent solutions, with neither colour 'signpost' appearing to be more effective than the other (Whibley *et al.*, 2006). Hybrid individuals can be observed at natural hybrid zones, showing intermediate flower colour phenotypes. However, these do not proliferate outside of *A. m. pseudomajus* and *A. m. striatum* contact zones, imlpying that they are generated through ongoing hybridisation (Field *et al.*, preliminary data).

In light of these observations, I hypothesise that flower colour in *A. m. pseudomajus* and *A. m. striatum* is a coadapted trait, arising through the interactions of at least seven loci from six distinct genomic regions. Colour loci together define the characteristic colour signposts that differentiate the two subspecies. For a colour pattern to be faithfully rendered, the correct alleles of all colour genes must be inherited. The substitution of any parental allele is sufficient to disrupt the coadapted colour phenotype, and thereby impart a substantial fitness penalty. These hybrid incompatibilities form the basis of the intrinsic epistatic barrier.

It is possible that the divergence between *A. m. pseudomajus* and *A. m. striatum* represents an example of reinforcement where, following secondary contact, selection against hybrid phenotypes can result in increased reproductive isolation. A similar barrier is hypothesised to exist between sympatric populations of *Phlox drummondii* and *Phlox cuspidata*. Here, two genes in the anthocyanin biosynthesis pathway show cis-regulatory variation between a dark red flowered morph of *P. drummondii*, and the light blue flowered *P. cuspidata* (Hopkins and Rausher, 2011). Each population has one dominant and one recessive variant meaning that, as with *A. m. pseudomajus* and *A. m. striatum*, hybrids show intermediate flower colours. While all interspecific hybrids show reduced fertility, dark red flowered *P. drummondii* show 66 % less interspecific hybridisation than *P. drummondii* morphs with light blue flowers (Levin, 1984). This suggests that magenta colour variation alone is sufficient to significantly reduce fitness. More recent evidence has suggested that reinforcement can proceed even where gene flow is ongoing, provided that the differentiated trait is strongly selected for (Roda *et al.*, 2017).

The observation that all six monophyletic regions identified within *A. m. pseudomajus* and *A. m. striatum* contain candidate colour genes is a compelling hint that colour alone may underlie the reproductive barrier. Flower colour in *A. m. pseudomajus* and *A. m. striatum* is mainly determined by the activity of two distinct but overlapping pigment biosynthesis pathways, downstream of the general flavonoid pathway. The anthocyanin pathway is complicated, with at least seven enzymes being required for anthocyanin biosynthesis in plants (Holton and Cornish, 1995). The aurone pathway, by comparison, is much simpler, consisting of two genes in *Antirrhinum* (Ono *et al.*, 2006). Because the aurone pathway is short, there are limited means by which mutation can act to change yellow pigment biosynthesis. If *FLA* and *AUN* represent true genetic barriers, then they together account for the entirety of the aurone pathway. A third proposed barrier locus, *SULF*, also acts through the aurone pathway by regulating *FLA* expression. Preliminary evidence from Desmond Bradley suggests that *CRE* may also influence yellow flower colour, although the mechanism of action remains to be characterised. Four genes, within two monophyletic islands, are predicted to regulate colour through the anthocyanin pathway; *RUB*, *ROS1*, *ROS2*, and *EL*.

Many anthocyanin-regulating genes have been characterised as major effect regulators of magenta patterning in *Antirrhinum*. The suggestion that only *ROS1*, *ROS2*, *EL*, and *RUB* show relevance in phenotypic divergence is surprising. Perhaps the most striking omission is *VENOSA* (*VE*), a gene encoding a MYB-like transcription factor that determines the patterning of the magenta anthocyanin pigments within a subset of epidermal cells overlying the veins, on the dorsal petals (Schwinn *et al.*, 2006). In a study of pollinator attraction, Shang *et al.* (2010) demonstrated that plants with dominant *VE* alleles (*i.e.* those which showed coloured veins) were more regularly visited by pollinators, if their flowers were otherwise pale coloured, or lacking anthocyanin. *VE* activity is unlikely to increase pollinator attractiveness of *A. m. pseudomajus* flowers, as the pattern is unlikely to stand out against the uniformly magenta background. However, magenta veins are expected to contrast well against the yellow coloured, acyanic flowers of *A. m. striatum*. Because *VE* is expected to have little effect on fitness in *A. m. pseudomajus*, *VE* may not be coadapted with other colour genes. If *VE* affects fitness within *A. m. striatum* populations only, then *VE* alleles should flow freely within *A. m. pseudomajus* populations, and therefore not underlie a genetic barrier. This could be tested by comparing π_W at the *VE* locus in *A. m. pseudomajus* and *A. m. striatum*. If *VE* confers additional fitness in *A. m. striatum*, then it may show reduced π_W, consistent with a selective sweep.

Support for genetic barriers and hybridisation in ancient growth habit divergence

For a genetic barrier to arise through differential adaptation, I hypothesised that three requirements must be met. Firstly, the underlying loci must represent a better solution for a given environment. Secondly, different adaptive solutions must be innovated for different environments. Finally, mixed strategies could work in an intermediate environment. My analyses of growth habit stop short of characterising functional loci, making it difficult to draw conclusions as to whether identified monophyletic islands are involved in adaptive divergence. Only one gene involved in growth habit divergence in *Antirrhinum*, *Hairy*, has been studied in detail (Tan *et al.*, 2020). This locus did fall within one of the of the 427 identified monophyletic regions, providing tantalising evidence that other such loci may be isolated. The characterisation of the genic landscape of other monophyletic islands therefore represents a future aim of high priority.

In focussing solely on the genic basis of monophyletic regions, it is easy to lose sight of other biological questions that can be addressed. Using a conservative method of size quantification, I observed a mean monophyletic island size of ~ 7 kb. This is likely to be an underestimate, casting doubt on recruitment of standing variation as a source of alleles for growth habit divergence. Even assuming a low recombination rate of 0.3 cM/Mb, recombination should break haplotypes down to < 3 kb within *ca.* 1,000 years. The amount of genomic divergence observed between alpine and ruderal populations is much too high for such a recent divergence to be considered. I hypothesise that monophyletic islands arose through allele sharing (introgression), or they represent historic barriers to gene flow, possibly originating through differential adaptation. Differentiating between these scenarios requires testing whether the underlying genetic variation is sporadically distributed throughout wild populations (implying an introgressive origin), or whether it is consistent. If growth habit divergence occurs mostly through introgression, then the number of identified monophyletic islands may vary in analyses of different populations. It should be noted that these scenarios are not mutually exclusive. A single, ancient divergence event between alpine and ruderal species may have been driven by a subset of genetic barriers which enabled populations to colonise different habitats. Populations may have then accumulated different adaptive alleles in isolation, and shared them during periods of contact.

Monophyletic regions showed elevated D_{XY} compared to the genomic mean, suggesting that they may reflect barriers to gene flow. However, mean SRB of trees at monophyletic regions was not significantly greater than trees showing the most common genomic topology. This suggests that more than one topology accounts for significant between-population divergence.

The most common topology groups *A. molle* with ruderal populations, suggesting that much gene flow has taken place. Despite this, *A. molle* is not phenotypically intermediate between alpine and ruderal – it retains striking resemblance to other alpine species (Rothmaler, 1956, Webb, 1971, Sutton, 1988). The simplest explanation for this is that growth habit identity is maintained by a smaller subset of genomic loci. An alternative hypothesis is that the most common genomic topology might reflect 'weak' growth habit loci. I define a weak locus as one that has a very small effect on adaptive fitness. Weak adaptive loci conferring slight fitness benefits may accumulate in isolation, or through allele sharing with populations showing the same growth habit. However, their effects might be so weak that they are readily displaced by gene flow with populations showing different growth habits. The weak habit model predicts that *A. molle* has undergone gene flow with ruderal species much more recently than other alpine species have. Contact between *A. molle* and ruderal populations is unlikely to have taken place at 'extremes' of their respective environmental ranges. It is possible that weakly adaptive variation is specifically adaptive within extreme alpine environments, but selectively neutral otherwise.

Many SNF candidates IRs exist, but their biological relevance cannot be tested *in silico*

SNF, a novel type of neofunctionalisation where coding sequence changes give rise to regulatory sRNA loci, may be an important mechanism underlying phenotypic divergence. Testing the hypothesis of SNF was beyond the scope of this computational analysis. Even if identified SNF candidate IRs were near identical to *SULF*, their functional relevance could not be tested without carrying out genetic analyses. However, the results provide insights which can inform how SNF loci are considered in future studies. Firstly, while up to 645 IRs were *SULF*-like in terms of their sRNA profiles, possible, targets, and presence / absence relationships between species, this only represented 0.5 % of all IRs tested. Assuming that many of these will be false positives or miRNAs, SNF candidate loci appear to be fairly rare within the genome. The main reason for this appears to be the ubiquity of IRs showing mode-24 nt mapped sRNAs. Secondly, while BLASTX searches provide limited functional information, they can provide hints as to the type of protein families that might be targeted by SNF elements. These analyses showed a slight enrichment (albeit from a very small sample size) for BLASTX hits against developmentally relevant F-box and auxin responsive domains. Finally, while grouping tree scans are likely ineffective for direct identification of SNF candidate IRs, results represent a useful resource for testing externally identified loci. A drawback to this comparison is that IRs must be present within the reference genome used to carry out the grouping tree scan. A possible workaround is to map not the IR (which may be present or absent), but its surrounding genomic regions, which should be relatively consistent in any *Antirrhinum* species. A long term aim of the grouping tree scan work could be to define a set of genomic regions, comparable between assemblies, that show monophyletic signatures for traits of interest.

Genetic barriers are underpinned by dispersed genomic loci

By considering the genomic context of monophyletic regions, it is possible to develop an idea of what barrier gene signatures generally look like. In *A. m. pseudomajus* and *A. m. striatum,* up to 0.47 % of the genome comprised monophyletic regions, and was implicated in the genetic barrier. This represents 2.38 Mb of sequence, containing a total of 968 predicted CDSs. The average gene density in *Antirrhinum majus* is estimated at 1 gene / 15.5 kb (Li *et al.*, 2019), meaning that a 2.38 Mb region would be expected to harbour 154 genes on average. This suggests that monophyletic regions are enriched for CDSs. However, the genomic average gene density is expected to be low given that over 50 % of the *A. majus* genome consists of repetitive sequence. Many coding sequences will not be expressed, and those that are will not necessarily be involved in phenotypic divergence. My analysis of Annabel Whibley's RNAseq DE data from *A. m. pseudomajus* and *A. m. striatum* reported 29 DE genes over all six monophyletic regions.

In the multispecies growth-habit-based comparisons*,* estimates from 50 kb windows suggested that monophyletic regions encompassed 15.4 Mb, or 3.1 % of the genome. These regions contained 1,219 CDSs. This is greater than the 995 genes expected based on average gene density, but represents less of an enrichment than in *A. m. pseudomajus* and *A. m. striatum*. The absence of RNAseq data for these populations means that any or all of these CDSs may be involved in divergence around growth habit.

DE analysis quantifies differences in mRNA transcript levels, and therefore doesn't detect changes in non-mRNA-producing loci. The characterisation of *SULF* as a locus that likely underlies flower colour divergence (Bradley *et al.*, 2017) demonstrates the importance of thinking beyond protein-coding genes when characterising genetic barriers.

In both grouping tree scan experiments, identified monophyletic regions were spread throughout the genome, and did not show strong biases to any particular chromosomal regions. This suggests that the approach is robust to variation in recombination rate across chromosomes. Monophyletic island size can provide an estimate of local recombination rate. Larger islands are likely to reflect regions where recombination rate is low, and genetic barrier loci segregate within large haplotype blocks. In practice, this means that large monophyletic islands are more likely to contain additional genes that do not directly contribute to observed patterns of divergence. Where recombination rate is higher, candidate barrier genes are easier to identify. This is most notable at the *RUB* region; only one gene within the 50 kb monophyletic island was differentially expressed.

In these analyses, larger islands tended to be localised towards the centre of the chromosome, and smaller islands towards the edges. This loosely corresponds to the observed patterns of recombination rate across *Antirrhinum* chromosomes (Li *et al.*, 2019). Centrally located islands were estimated to be larger than those within outer regions. However, within this work I have only directly measured the sizes of 21 monophyletic islands. When estimating the size of the remaining 412 monophyletic islands, I reported the minimum island size, which reflected the width of the island if all overlapping 1 kb monophyletic windows were directly adjacent. This approach only operates within the 50 kb boundaries of the originally identified monophyletic region, meaning that it cannot detect instances where a monophyletic island extends beyond the region. Minimum island size will therefore almost always underestimate the width of a monophyletic island.

To best estimate whether island size correlates with chromosomal position and genomic recombination rate, the size of the monophyletic island should be quantified within all monophyletic regions. The approach to measuring islands presented here involves defining the start and end of the island by-eye. This is neither objective, nor easy to replicate *en masse*. An approach is needed that determines how far 1 kb monophyletic windows extend out of the 50 kb monophyletic region, whilst allowing for reasonably sized gaps that might reflect poor read depth or sequencing errors. The Saguaro tool addresses a similar problem, detection of phylogenetic boundaries within genomes in the absence of *a priori* assumptions, through use of Hidden Markov Models (Zamani *et al.*, 2013).

Genetic barriers show varied signatures of diversity, consistent with distinct origins

In carrying out this work, I have demonstrated a range of parameters that can be used to make inferences about the evolutionary history of a given monophyletic island. Of these, D_{XY}, π_W, D, and F_{ST} have been previously defined and extensively utilised. SRB, defined here, is equivalent to d_f, as defined by Hey (1991). This measures the number of allelic differences between populations, excluding those which also differ within populations. For two populations, X and Y, d_f is equal to $D_{XY} - \pi_X$ or $D_{XY} - \pi_y$, where π is within-population diversity (*i.e.* π_W). In deriving SRB, I effectively calculate d_f using π_W for each population, and choose the smallest (non-zero) value. The derivation of D is very similar, except that the mean of π_X and π_Y is subtracted. d_f, SRB and D are relative measures of divergence, because they depend on π_W. A large part in my decision to use SRB was its descriptive nature; the shortest root branch is easily observable on any tree. However, D has a more intuitive biological relevance, being equal to the number of differences between the populations since they split.

It has been shown that genes underpinning the same genetic barriers can show strikingly different patterns of allelic diversity. The *ROS-EL* region, in comparisons between *A. m. pseudomajus* and *A. m. striatum*, revealed that genes can show distinct divergence landscapes even if they are tightly linked. This region showed a broad F_{ST} peak, which has been previously characterised by Tavares *et al.* (2018) as corresponding to the location of the *ROS1* and *ROS2* genes, and another corresponding to *EL*. However, only the *EL* peak showed a coincident peak in D, consistent with elevated D_{XY}. This demonstrated that the major component of elevated F_{ST} at *ROS1* and *ROS2* was reduced π_W, consistent with an historical selective sweep. A similar signature was seen at the *FLA* region, with high F_{ST} but only slightly elevated D across most of the monophyletic island. Preliminary work by Bradley *et al.* has demonstrated that the monophyletic island, and coincident drop in D, reflects the boundary between the *FLA* gene CDS, and upstream promoter. They have also observed that, at the *A. m. pseudomajus* and *A. m.*

striatum hybrid zone at Planoles, a recombinant *FLA* allele comprising the promoter region of *A. m. striatum* and the CDS of *A. m. pseudomajus* has been detected. The sudden shift from monophyletic trees to polyphyletic trees across *FLA* may reflect the fact that the recombinant *FLA* allele exists at high frequency within the hybrid zone YP1 population, which therefore groups within the *A. m. pseudomajus* clade. Characterisation of this recombinant is ongoing (Bradley *et al.*, manuscript in preparation).

The grouping tree scan – advantages and alternatives

To date, no published studies have utilised D_{XY} tree comparisons to infer barrier loci across whole genomes, but similar methodologies exist for characterising genomic divergence across populations. Where the grouping tree scan generates a tree for each genomic region, sliding window Principal Component Analysis (PCA) computes the most descriptive dimensions of the SNP data, which can be summarised using eigenvectors (see *e.g.* Jay *et al.*, 2021). Similarly, Saguaro (Zamani *et al.*, 2013) achieves multiple comparisons by using a Hidden Markov Model to infer genomic regions showing shared phylogenies. BayeScan (Fischer *et al.*, 2011) uses Bayesian inference to detect F_{ST} outliers between populations, and thereby infer loci under natural selection. BayeScan models selection by estimating a population-component (which is shared by all loci) and loci-specific components (which are shared by all populations) from F_{ST} distributions. In this sense, population comparisons are "multi-pairwise" in the same way as other measures described here.

The value of the grouping tree scan is in its combining of D_{XY} analysis, hierarchical clustering, and whole genome scans. D_{XY} enables the detection of genetic divergence independently of within-population diversity, meaning that it can be used to infer barriers to gene flow directly from allele frequency data. BayeScan, like Twisst, utilises F_{ST}, which conflates within- and between-population diversity, and PCA derives the components of diversity in a manner that is naïve of this distinction. Saguaro utilises absolute genetic distances, but the derived data structures are not designed to be biologically intuitive. The typical drawback of using D_{XY} is that it that it is highly heterogeneous across the whole genome, and therefore difficult to interpret in pairwise analyses. Hierarchical clustering accounts for this weakness by facilitating multi-pairwise analysis of D_{XY}, which helps to reduce sporadic noise whilst increasing genuine signals. UPGMA trees have several properties that make them useful for exploring large datasets. They can be clustered based on topology, which facilitates comparisons across large numbers of genomic regions. Simple properties, such as tree height and *SRB*, intuitively reflect the underlying biological phenomena. The derivation of D_{XY} from allele frequencies, and UPGMA trees from D_{XY} distributions, is mathematically straightforward. The relative effectiveness of the aforementioned approaches will ultimately depend on the biological question, the nature of the available data, and the expertise of the researcher. Indeed, characterising the differences between results from different tools prove complementary for understanding the nature of the allelic divergence at loci of interest.

A note on library barcoding

The grouping tree scan utilises pool-seq as a cost-effective means of capturing as much population allelic variation as possible. As with individual sequencing, DNA is extracted from each sample in turn. However, by pooling equivalent quantities of DNA from all sample, SNPs can be identified from all individuals in one sequencing run. In instances where read coverage is high (*e.g.* when a good quality reference genome is available), pool-seq is more efficient for SNP calling than individual sequencing (Futschik and Schlötterer, 2010). If the objective of the pool-seq experiment is to identify SNPs across populations, then the sample of origin of individual sequencing reads might not be important. However, once this information is lost, it becomes significantly more difficult to link genotypic data to individual phenotypes. It also hinders haplotype identification, necessitating the use of predictive bioinformatic tools to deduce individual haplotype sequences based on allele frequencies (Wong *et al.*, 2018). Recovery of haplotypes is important for identifying structural variation, which has been frequently implicated in population variation (Marroni, Pinosio, and Morgante, 2014, Ruggieri *et al.*, 2022, Hollox, Zuccherato, and Tucci, 2022). Multiplex sequencing pipelines, which aim to sequence many distinct libraries in a single sequencing run, typically include a barcoding step, where an identifying nucleotide sequence is ligated to all reads within a given library. By barcoding individual libraries prior to sequencing, it is possible to link genetic variation to multiple accessions of origin, even if reads are subsequently pooled. Because this requires the generation of multiple libraries, it is more expensive, although advances to technology and infrastructure have rendered library preparation increasingly cost effective (Head *et al.*, 2014).

Continuing development of the grouping tree scan approach

The grouping tree scan represents a promising extension of the wider genome scan toolkit, having been successfully applied to interpret otherwise noisy genomic landscapes of D_{XY}. A limitation of the grouping tree scan in its current form is that initial identification of monophyletic regions is tied to windowed averages of D_{XY}, and may lack the sensitivity to detect smaller monophyletic islands. Therefore, in the further development of the grouping tree scan approach, I propose that scans should be carried out using trees from individual genomic sites, rather than window averages. Methodologically, this is mostly identical to the scan as described here. The same mapping and data generation steps should be carried out. However, instead of carrying out a full SlidingWindows analysis, only site statistics should be generated. Two statistics, D_{XY} and π_W, should be calculated for each genomic site. This is already done during the SlidingWindows analysis – the output site file can be used. Alternatively, a streamlined approach could be developed, producing a smaller data table which required less computational power to process. For each site, a D_{XY} distance matrix, an UPGMA tree, and a mean value of π_W should be generated, and stored. This is the minimum amount of data required for all downstream computations. This thesis has utilised two tree classification approaches – randomly seeded forest clustering using the cophenetic correlation coefficient, and root division analysis. Analyses carried out here suggest that each of these approaches is better suited to a different biological problem – both could be implemented for single site grouping tree scans. Ultimately, the aim of the approach is to plot the density of SNPs showing trees of interest (based on forest *SRB*, or user specified root division topologies).

Having generated single site tree data, many analyses become possible. To consider some of these, I shall consider how they might be applied to the study of growth habit. Having established all genomic sites that show monophyletic trees for alpine and ruderal growth habits, all predicted CDSs that overlap these sites could be extracted. This would facilitate the capture of the full complement of genes that might be involved in growth habit divergence. Additionally, chromosomal distributions of monophyletic SNPs could be visualised, revealing whether specific regions of the genome show enrichment or depletion. Approaches can be developed to measure the number and size of islands, which may inform about their evolutionary origins. This approach has the potential to facilitate thorough characterisation of genomic islands of divergence, at the highest possible resolution afforded by mapping-based approaches.

7: Conclusions

Chapter 3

Application of the grouping tree scan methodology has revealed that six monophyletic regions, comprising 0.47 % of the genome, underpin genomic divergence between *A. m. pseudomajus* and *A. m. striatum*. This is in contrast to the whole genome tree, which is polyphyletic for both subspecies and consistent with extensive gene flow. Each monophyletic region contained at least one differentially expressed candidate colour gene, all of which have been directly or indirectly implicated in controlling flower colour patterns. Because the vast majority of genomic divergence was explained by flower colour, the heterozygote advantage hypothesis can be rejected. These analyses provide strong evidence that genetic barrier between *A. m. pseudomajus* and *A. m. striatum* is underpinned by colour genes only, and reflects an intrinsic epistatic barrier. Future experiments should aim to statistically analyse the proposed role of candidate colour genes in generating distinct flower colour phenotypes between species. Cline analysis should be carried out to confirm that uncharacterised loci show coincident clines, and are therefore likely to be coadapted.

Chapter 4

To test the hypothesis that divergence between alpine and ruderal species involved differentially adaptive barrier genes, genome scans were first carried out using F_{ST}, D_{XY}, and π_W. No clear divergence peaks were detectable, demonstrating that two-way comparisons were confounded by the uneven genomic divergence landscape. A grouping tree scan was carried out to classify subgenomic D_{XY} trees. Trees were classified into topological groups based on root division, which splits each tree into the two outermost clades and records taxa in each clade. The whole genome D_{XY} tree, and the most common subgenomic tree, gave polyphyletic groupings for growth habits, suggesting that gene flow has taken place between alpine and ruderal species. 427 subgenomic regions gave trees that were monophyletic for alpine and ruderal growth habits. Conservative estimates of the sizes of monophyletic islands at these regions suggest a mean island size of ~ 7 kb. This is much larger than predicted if growth habit divergence took place through recruitment of alleles from standing genetic variation. Of the 15 monophyletic islands that have been characterised, the largest was ~ 375 kb, which could be consistent with allele sharing through hybridisation, or barrier genes. Identified monophyletic regions may therefore underpin a differentially adapted genetic barrier. More work is required to characterise the genic basis of identified islands, and to better estimate the size of all islands.

Chapter 5

This chapter has outlined a pipeline for detecting *SULF*-like SNF candidate IRs, to test the hypothesis that SNF is a general phenomenon by which phenotypic diversity can arise. 645 SNF candidate IRs passed criteria 1-4, out of 131,755 tested. This shows that more SNF loci can be detected, but they constitute a minority of all genomic IRs. This candidate set is likely to include many false positives – more work needs to be done to exclude candidates with poor read mapping profiles. 34 candidates were selected for comparison to the NCBI nr database using BLASTX. This implied a role for SNF in targeting developmental genes, but few conclusions can be drawn based on predictions of protein similarity alone. Comparing SNF candidate IRs to the growth habit monophyletic region dataset from Chapter 4 showed that two IRs are located within monophyletic regions. However, one of these IRs showed very low sRNA depth, and may be an artifact. The other appeared to be present within the alpine *A. molle*, and the ruderal *A. majus* reference genome. This casts doubt on whether it is likely to be involved in growth habit divergence. To test the evolutionary significance of SNF, the candidate list derived here should be further filtered to retain only IRs showing valid sRNA distributions, and characterised target genes. The effects of these SNF candidates could then be tested using segregation analysis, where *Antirrhinum* species showing presence / absence for candidate IRs can be crossed, and phenotypes analysed in a hybrid F_2 population. Analyses should also be expanded beyond *Antirrhinum*. Using the SNF program, Leighton Folkes *et al*. have detected SNF candidate loci within the genomes of *Arabidopsis thaliana* and *Solanum lycopersicum* (Folkes *et al.*, manuscript in preparation). Even if *SULF* is a 'one-off' locus in *Antirrhinum*, other SNF loci should be detectable in different genera.

8: References

Aagaard, J.E., George, R.D., Fishman, L., MacCoss, M.J., and Swanson, W.J. (2013). Selection on Plant Male Function Genes Identifies Candidates for Reproductive Isolation of Yellow Monkeyflowers. PLOS Genet. 9, e1003965.

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife 4, e05005.

Akita, Y., Kitamura, S., Hase, Y., Narumi, I., Ishizaka, H., Kondo, E., Kameari, N., Nakayama, M., Tanikawa, N., Morita, Y., and Tanaka, A. (2011). Isolation and characterization of the fragrant cyclamen O-methyltransferase involved in flower coloration. Planta 234, 1127–1136.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.H., Spatafora, J.W., and Carrington, J.C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat. Genet.

Angert, A.L., and Schemske, D.W. (2005). The evolution of species' distributions: Reciprocal transplants across the elevation ranges of Mimulus cardinalis and M. lewish. Evolution. 59, 1671–1684.

Arendt, J., and Reznick, D. (2008). Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol. Evol. 23, 26–32.

Arkhipova, I.R., and Yushenova, I.A. (2019). Giant Transposons in Eukaryotes: Is Bigger Better? Genome Biol. Evol. 11, 906–918.

Axtell, M.J. (2013). ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA 19, 740–751.

Barton, N.H., and Hewitt, G.M. (1985). Analysis of Hybrid Zones. Annu. Rev. Ecol. Syst. 16, 113–148.

Begun, D.J., and Aquadro, C.F. (1991). Molecular population genetics of the distal portion of the X chromosome in Drosophila: Evidence for genetic hitchhiking of the yellow-achaete region. Genetics 129, 1147–1158.

Borges, F., and Martienssen, R.A. (2015). The expanding world of small RNAs in plants. Nat. Rev. Mol. Cell Biol.

Bradley, D., Xu, P., Mohorianu, I.I., Whibley, A., Field, D., Tavares, H., Couchman, M., Copsey, L., Carpenter, R., Li, M., Li, Q., Xue, Y., Dalmay, T., and Coen, E. (2017). Evolution of flower color pattern through selection on regulatory small RNAs. Science (80-.). 358, 925–928.

Brien, M.N., Enciso-Romero, J., Lloyd, V.J., Curran, E.V., Parnell, A.J., Morochz, C., Salazar, P.A., Rastas, P., Zinn, T., and Nadeau, N.J. (2022). The genetic basis of structural colour variation in mimetic Heliconius butterflies. Phil. Trans. R. Soc. B. 377.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinformatics, 3(1), lqaa108.

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics and Bioinformatics, 2(2), 1–14.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. 38, 5825–5829.

Carpenter, R., Martin, C., and Coen, E.S. (1987). Comparison of genetic behaviour of the transposable element Tam3 at two unlinked pigment loci in Antirrhinum majus. Mol. Gen. Genet. 207, 82–89.

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M., Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., and Schatz, M.C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054.

Chouteau, M., Arias, M., and Joron, M. (2016). Warning signals are under positive frequency dependent selection in nature. Proc. Natl. Acad. Sci. U. S. A. 113, 2164–2169.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibé, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., Larzul, C., Laville, E., Meish, F., Milenkovic, D., Tobin, J., Charlier, C., and Georges, M. (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat. Genet.

Coen, E.S., Carpenter, R., and Martin, C. (1986). Transposable elements generate novel spatial patterns of gene expression in Antirrhinum majus. Cell 47, 285–296.

Cooley, A.M., and Willis, J.H. (2009). Genetic divergence causes parallel evolution of flower color in Chilean Mimulus. New Phytol. 183, 729–739.

Coyne, J.A., Orr, H.A. (2004). Speciation (Sinauer associates Sunderland, MA).

Crow, J. F., and Kimura, M. (1970). An Introduction to Genetics Theory.

Cruickshank, T.E., and Hahn, M.W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23, 3133–3157.

Csilléry, K., Rodríguez-Verdugo, A., Rellstab, C., and Guillaume, F. (2018). Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. Mol. Ecol. 27, 606–612.

Cui, J., You, C., and Chen, X. (2017). The evolution of microRNAs in plants. Curr. Opin. Plant Biol.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10, giab008.

Davies, K.M., Marshall, G.B., Bradley, J.M., Schwinn, K.E., Bloor, S.J., Winefield, C.S., and Martin, C.R. (2006). Characterisation of aurone biosynthesis in Antirrhinum majus. Physiol. Plant. 128, 593–603.

Debernardi, J.M., Lin, H., Chuck, G., Faris, J.D., and Dubcovsky, J. (2017). microRNA172 plays a crucial role in wheat spike morphogenesis and grain threshability. Development.

Dixon, G., Kitano, J., and Kirkpatrick, M. (2019). The Origin of a New Sex Chromosome by Introgression between Two Stickleback Fishes. Mol. Biol. Evol. 36, 28–38.

Dobzhansky, T.H. (1950). Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudoobscura. Genetics 35, 288–302.

Du, H., Wu, J., Ji, K.-X., Zeng, Q.-Y., Bhuiya, M.-W., Su, S., Shu, Q.-Y., Ren, H.-X., Liu, Z.-A., and Wang, L.-S. (2015). Methylation mediated by an anthocyanin, O-methyltransferase, is involved in purple flower coloration in Paeonia. J. Exp. Bot. 66, 6563–6577.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science (80-.). 356, 92–95.

Durán-Castillo, M., Hudson, A., Wilson, Y., Field, D.L., and Twyford, A.D. (2022). A phylogeny of Antirrhinum reveals parallel evolution of alpine morphology. New Phytol. 233, 1426–1439.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 3, 99–101.

Eamens, A., Wang, M.B., Smith, N.A., and Waterhouse, P.M. (2008). RNA Silencing in Plants: Yesterday, Today, and Tomorrow. Plant Physiol.

Ellis, T.J., and Field, D.L. (2016). Repeated gains in yellow and anthocyanin pigmentation in flower colour transitions in the Antirrhineae. Ann. Bot.

Favre, A., Widmer, A., and Karrenberg, S. (2017). Differential adaptation drives ecological speciation in campions (Silene): evidence from a multi-site transplant experiment. New Phytol. 213, 1487–1499.

Feng, X., Wilson, Y., Bowers, J., Kennaway, R., Bangham, A., Hannah, A., Coen, E., and Hudson, A. (2009). Evolution of allometry in Antirrhinum. Plant Cell 21, 2999–3007.

Fischer, M.C., Foll, M., Excoffier, L., and Heckel, G. (2011). Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (Microtus arvalis). Mol. Ecol. 20, 1450–1462.

Foote, A.D. (2018). Sympatric Speciation in the Genomic Era. Trends Ecol. Evol. 33, 85–95.

Franks, S.J., Sim, S., and Weis, A.E. (2007). Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. Proc. Natl. Acad. Sci. 104, 1278–1282.

Fuentes, R.R., De Ridder, D., Van Dijk, A.D.J., and Peters, S.A. (2022). Domestication Shapes Recombination Patterns in Tomato. Mol. Biol. Evol. 39.

Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186, 207–218.

Gramlich, S., Liu, X., Favre, A., Buerkle, C.A., and Karrenberg, S. (2022). A polygenic architecture with habitat-dependent effects underlies ecological differentiation in Silene. New Phytol. 235, 1641–1652.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34, D140–D144.

Guan, D., Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., Durbin, R., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9), 2896–2898.

Hamilton, A.J., and Baulcombe, D.C. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. Science (80-.). 286, 950–952.

Hannon, G.J., Rivas, F. V., Murchison, E.P., and Steitz, J.A. (2006). The expanding universe of noncoding RNAs. In Cold Spring Harbor Symposia on Quantitative Biology, p.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56, 61–64, 66, 68, passim.

Hey, J. (1991). The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics 128, 831–840.

Hollox, E.J., Zuccherato, L.W., and Tucci, S. (2022). Genome structural variation in human evolution. Trends Genet. 38, 45–58.

Holton, T.A., and Cornish, E.C. (1995). Genetics and Biochemistry of Anthocyanin Biosynthesis. Plant Cell 7, 1071–1083.

Hopkins, R., and Rausher, M.D. (2011). Identification of two genes causing reinforcement in the Texas wildflower Phlox drummondii. Nature 469, 411–414.

Jaworski, C.C., Andalo, C., Raynaud, C., Simon, V., Thébaud, C., and Chave, J. (2015). The influence of prior learning experience on pollinator choice: An experiment using bumblebees on two wild floral types of Antirrhinum majus. PLoS One.

Jaworski, C.C., Thébaud, C., and Chave, J. (2016). Dynamics and persistence in a metacommunity centred on the plant Antirrhinum majus: Theoretical predictions and an empirical test. J. Ecol.

Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., and Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nat. Genet. 53, 288–293.

Jensen, M.A., Charlesworth, B., and Kreitman, M. (2002). Patterns of genetic variation at a chromosome 4 locus of Drosophila melanogaster and D. simulans. Genetics 160, 493–507.

Jia, L., Li, Y., Huang, F., Jiang, Y., Li, H., Wang, Z., Chen, T., Li, J., Zhang, Z., and Yao, W. (2022). LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes. Nucleic Acids Res. 50, D174–D182.

Jiggins, C.D., and Martin, S.H. (2017). Glittering gold and the quest for Isla de Muerta. J. Evol. Biol. 30, 1509–1511.

Karrenberg, S., Liu, X., Hallander, E., Favre, A., Herforth-Rahmé, J., and Widmer, A. (2018). Ecological divergence plays an important role in strong but complex reproductive isolation in campions (Silene). Evolution. 73, 245–261.

Kaufmann, K., and Airoldi, C.A. (2018). Master Regulatory Transcription Factors in Plant Development: A Blooming Perspective BT - Plant Transcription Factors: Methods and Protocols. N. Yamaguchi, ed. (New York, NY: Springer New York), pp. 3–22.

Khimoun, A., Burrus, M., Andalo, C., Liu, Z.L., Vicédo-Cazettes, C., Thébaud, C., and Pujol, B. (2011). Locally asymmetric introgressions between subspecies suggest circular range expansion at the Antirrhinum majus global scale. J. Evol. Biol. 24, 1433–1441.

Kincaid, P. (1986). The Rule of the Road: An International Guide to History and Practice. (Bloomsbury Publishing Plc.)

Knauss, S., Rohrmeier, T., and Lehle, L. (2003). The Auxin-induced Maize Gene ZmSAUR2 Encodes a Short-lived Nuclear Protein Expressed in Elongating Tissues. J. Biol. Chem. 278, 23936–23943.

Knief, U., Bossu, C.M., Saino, N., Hansson, B., Poelstra, J., Vijay, N., Weissensteiner, M., and Wolf, J.B.W. (2019). Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. Nat. Ecol. Evol. 3, 570–576.

Kofler, R., Pandey, R.V., and Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Research, 27(5), 722–736.

Krueger, F. (2015). Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Babraham Institute.

Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays. (2013) Sep;35(9):780-6. doi: 10.1002/bies.201300014. Epub 2013 Jul 9. PMID: 23836388; PMCID: PMC3849385.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.

Lankinen, Å., and Karlsson Green, K. (2015). Using theories of sexual selection and sexual conflict to improve our understanding of plant ecology and evolution. AoB Plants 7.

Láruson, Á.J., and Reed, F.A. (2016). Stability of underdominant genetic polymorphisms in population networks. J. Theor. Biol. 390, 156–163.

Lee, C.H., and Carroll, B.J. (2018). Evolution and Diversification of Small RNA Pathways in Flowering Plants. Plant Cell Physiol. 59, 2169–2187.

Leonard, A.S., Brent, J., Papaj, D.R., and Dornhaus, A. (2013). Floral Nectar Guide Patterns Discourage Nectar Robbing by Bumble Bees. PLoS One.

Levin, D.A. (1984). Inbreeding Depression and Proximity-Dependent Crossing Success in Phlox drummondii. Evolution (N. Y). 38, 116–127.

Lewsey, M.G., Hardcastle, T.J., Melnyk, C.W., Molnar, A., Valli, A., Urich, M.A., Nery, J.R., Baulcombe, D.C., and Ecker, J.R. (2016). Mobile small RNAs regulate genome-wide DNA methylation. Proc. Natl. Acad. Sci. U. S. A. 113, E801-10.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574.

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

Li, M., Zhang, D., Gao, Q., Luo, Y., Zhang, H., Ma, B., Chen, C., Whibley, A., Zhang, Y., Cao, Y., Li, Q., Guo, H., Li, J., Song, Y., Zhang, Y., Copsey, L., Li, Y., Li, X., Qi, M., Wang, J., Chen, Y., Wang, D., Zhao, J., Liu, G., Wu, B., Yu, L., Xu, C., Li, J., Zhao, S., Zhang, Y., Hu, S., Liang, C., Yin, Y., Coen, E., and Xue, Y. (2019). Genome structure and evolution of Antirrhinum majus L. Nat. Plants.

Lincoln, A., Cooper, J., and Loovers, J.P.L. (2020). Arctic: Culture and Climate (Thames and Hudson).

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

Lunau, K., Wacht, S., and Chittka, L. (1996). Colour choices of naive bumble bees and their implications for colour perception. J. Comp. Physiol. A 178, 477–489.

Luo, P., Ning, G., Wang, Z., Shen, Y., Jin, H., Li, P., Huang, S., Zhao, J., and Bao, M. (2016). Disequilibrium of Flavonol Synthase and Dihydroflavonol-4-Reductase Expression Associated Tightly to White vs. Red Color Flower Formation in Plants. Front. Plant Sci. 6.

Ma, X., Liu, C., Gu, L., Mo, B., Cao, X., and Chen, X. (2018). TarHunter, a tool for predicting conserved microRNA targets and target mimics in plants. Bioinformatics 34, 1574–1576.

Ma, Z., and Zhang, X. (2018). Actions of plant Argonautes: predictable or unpredictable? Curr. Opin. Plant Biol.

Mackin, C.R., Peña, J.F., Blanco, M.A., Balfour, N.J., and Castellanos, M.C. (2021). Rapid evolution of a floral trait following acquisition of novel pollinators. J. Ecol. 109, 2234–2246.

Mallet, J., and Barton, N.H. (1989). Strong natural selection in a warning-color hybrid zone. Evolution (N. Y). 43, 421–431.

Mallory, A.C., Bartel, D.P., and Bartel, B. (2005). MicroRNA-Directed Regulation of Arabidopsis AUXIN RESPONSE FACTOR17 Is Essential for Proper Development and Modulates Expression of Early Auxin Response Genes. Plant Cell 17, 1360–1375.

Marques, D.A., Meier, J.I., and Seehausen, O. (2019). A Combinatorial View on Speciation and Adaptive Radiation. Trends Ecol. Evol. 34, 531–544.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? Curr. Opin. Plant Biol. 18, 31–36.

Martin, C., Prescott, A., Mackay, S., Bartlett, J., and Vrijlandt, E. (1991). Control of anthocyanin biosynthesis in flowers of Antirrhinum majus. Plant J. 1, 37–49.

Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins, C.D., Kronforst, M.R., Long, A.D., McMillan, W.O., and Reed, R. (2012). Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. Proc. Natl. Acad. Sci. U. S. A. 109, 12632–12637.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C.D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23, 1817–1828.

Martin, S.H., and Van Belleghem, S.M. (2017). Exploring Evolutionary Relationships Across the Genome Using Topology Weighting. Genetics 206, 429–438.

McGaugh, S.E., Heil, C.S.S., Manzano-Winkler, B., Loewe, L., Goldstein, S., Himmel, T.L., and Noor, M.A.F. (2012). Recombination Modulates How Selection Affects Linked Sites in Drosophila. PLOS Biol. 10, e1001422.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The

Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.

Maechler M., Rousseeuw P., Struyf A., Hubert M., and Hornik K. (2022). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4

McClure, B.A., and Guilfoyle, T. (1989). Rapid Redistribution of Auxin-Rregulated RNAs During Gravitropism. Science (80-). 243, 91–93.

Moore, J.C., and Pannell, J.R. (2011). Sexual selection in plants. Curr. Biol. 21, R176–R182.

Moore, R.C., and Purugganan, M.D. (2005). The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol.

Morgado, L., and Johannes, F. (2017). Computational tools for plant small RNA detection and categorization. Brief. Bioinform.

Morris, J., Navarro, N., Rastas, P., Rawlins, L.D., Sammy, J., Mallet, J., and Dasmahapatra, K.K. (2019). The genetic architecture of adaptation: convergence and pleiotropy in Heliconius wing pattern evolution. Heredity (Edinb). 123, 138–152.

Nadeau, N.J., Pardo-Diaz, C., Whibley, A., Supple, M.A., Saenko, S. V., Wallbank, R.W.R., Wu, G.C., Maroja, L., Ferguson, L., Hanly, J.J., Hines, H., Salazar, C., Merrill, R.M., Dowling, A.J., Ffrench-Constant, R.H., Llaurens, V., Joron, M., McMillan, W.O., and Jiggins, C.D.. (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. Nature 534, 106–110.

Nakayama, T., Yonekura-Sakakibara, K., Sato, T., Kikuchi, S., Fukui, Y., Fukuchi-Mizutani, M., Ueda, T., Nakao, M., Tanaka, Y., Kusumi, T., and Nishino, T. (2000).

Aureusidin Synthase: A Polyphenol Oxidase Homolog Responsible for Flower Coloration. Science (80-). 290, 1163–1166.

Nakayama, T. (2022). Biochemistry and regulation of aurone biosynthesis. Biosci. Biotechnol. Biochem. 86, 557–573.

Nei, M., and Roychoudhury, A.K. (1973). Probability of Fixation and Mean Fixation Time of an Overdominant Mutation. Genetics 74, 371–380.

Nei, M., and Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U. S. A. 76, 5269–5273.

Nei, M. (1987). Molecular Evolutionary Genetics.

Newman, T.C., Ohme-Takagi, M., Taylor, C.B., and Green, P.J. (1993). DST sequences, highly conserved among plant SAUR genes, target reporter transcripts for rapid decay in tobacco. Plant Cell 5, 701–714.

Nosil, P., Funk, D.J., and Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18, 375–402.

Nozawa, M., Miura, S., and Nei, M. (2012). Origins and evolution of microRNA genes in plant species. Genome Biol. Evol.

Ohno, S. (1970). Evolution by Gene Duplication. Springer-Verlag, New York, 1970.

Okitsu, N., Mizuno, T., Matsui, K., Choi, S.H., and Tanaka, Y. (2018). Molecular cloning of flavonoid biosynthetic genes and biochemical characterization of anthocyanin o-methyltransferase of Nemophila menziesii Hook. and Arn. Plant Biotechnol. 35, 9–16.

Ono, E., Fukuchi-Mizutani, M., Nakamura, N., Fukui, Y., Yonekura-Sakakibara, K., Yamaguchi, M., Nakayama, T., Tanaka, T., Kusumi, T., and Tanaka, Y. (2006). Yellow flowers generated by expression of the aurone biosynthetic pathway. Proc. Natl. Acad. Sci.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research, 46(21), e126. https://doi.org/10.1093/nar/gky730

Panchy, N., Lehti-Shiu, M.D., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. Plant Physiol.

Panhuis, T.M., Butlin, R., Zuk, M., and Tregenza, T. (2001). Sexual selection and speciation. Trends Ecol. Evol. 16, 364–371.

Peñalba, J. V., Peters, J.L., and Joseph, L. (2022). Sustained plumage divergence despite weak genomic differentiation and broad sympatry in sister species of Australian woodswallows (Artamus spp.). Mol. Ecol. 5060–5073.

Petrov, D.A., and Hartl, D.L. (2000). Pseudogene evolution and natural selection for a compact genome. J. Hered. 91, 221–227.

Ponnikas, S., Sigeman, H., Lundberg, M., and Hansson, B. (2022). Extreme variation in recombination rate and genetic diversity along the Sylvioidea neo-sex chromosome. Mol. Ecol. 31, 3566–3583.

Qiao, H., Wang, F., Zhao, L., Zhou, J., Lai, Z., Zhang, Y., Robbins, T.P., and Xue, Y. (2004). The F-Box Protein AhSLF-S2 Controls the Pollen Function of S-RNase–Based Self-Incompatibility. Plant Cell 16, 2307–2322.

Ramachandran, S., Hiratsuka, K., and Chua, N.-H. (1994). Transcription factors in plant growth and development. Curr. Opin. Genet. Dev. 4, 642–646.

Ramsey, J., Bradshaw, H.D., and Schemske, D.W. (2003). Components of reproductive isolation between the monkeyflowers Mimulus lewisii and M. cardinalis (Phrymaceae). Evolution (N. Y). 57, 1520–1534.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, 11(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Ravinet, M., Faria, R., Butlin, R.K., Galindo, J., Bierne, N., Rafajlović, M., Noor, M.A.F., Mehlig, B., and Westram, A.M. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J. Evol. Biol. 30, 1450–1477.

Reed, R. D., Papa, R., Martin, A., Hines, H. M., Counterman, B. A., Pardo-Diaz, C., Jiggins, C. D., Chamberlain, N. L., Kronforst, M. R., Chen, R., Halder, G., Nijhout, H. F., and McMillan, W. O. (2011). Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. Science, 333(6046), 1137-1141. https://doi.org/10.1126/science.1208227

Reverté, S., Retana, J., Gómez, J.M., and Bosch, J. (2016). Pollinators show flower colour preferences but flowers with similar colours do not attract similar pollinators. Ann. Bot.

Ringbauer, H., Kolesnikov, A., Field, D.L., and Barton, N.H. (2018). Estimating Barriers to Gene Flow from Distorted Isolation-by-Distance Patterns. Genetics 208, 1231–1245.

Roda, F., Mendes, F.K., Hahn, M.W., and Hopkins, R. (2017). Genomic evidence of gene flow during reinforcement in Texas Phlox. Mol. Ecol. 26, 2317–2330.

Rokas, A., King, N., Finnerty, J., and Carroll, S.B. (2003). Conflicting phylogenetic signals at the base of the metazoan tree. Evol. Dev. 5, 346–359.

Rothmaler, W. (1956). Taxonomische Monographie de Gattung Antirrhinum. (Berlin: Academie-Verlag).

Ruggieri, A.A., Livraghi, L., Lewis, J.J., Evans, E., Cicconardi, F., Hebberecht, L., Ortiz-Ruiz, Y., Montgomery, S.H., Ghezzi, A., Rodriguez-Martinez, J.A., Jiggins, C.D., McMillan, W.O., Counterman, B.A., Papa, R., and Van Belleghem, S.M. (2022). A butterfly pan-genome reveals that a large amount of structural variation underlies the evolution of chromatin accessibility. Genome Res. 32, 1862–1875.

Schemske, D.W., and Bradshaw, H.D. (1999). Pollinator preference and the evolution of floral traits in monkeyflowers (Mimulus). Proc. Natl. Acad. Sci. 96, 11910–11915.

Schwarz-Sommer, Z., Davies, B., and Hudson, A. (2003). An everlasting pioneer: The story ofantirrhinum research. Nat. Rev. Genet.

Schwinn, K., Venail, J., Shang, Y., Mackay, S., Alm, V., Butelli, E., Oyama, R., Bailey, P., Davies, K., and Martin, C. (2006). A Small Family of MYB-Regulatory Genes Controls Floral Pigmentation Intensity and Patterning in the Genus Antirrhinum. The Plant Cell.

Seehausen, O., Butlin, R.K., Keller, I., Wagner, C.E., Boughman, J.W., Hohenlohe, P.A., Peichel, C.L., and Saetre, G. (2014). Genomics and the origin of species. Nat. Publ. Gr. 15.

Shang, Y., Venail, J., Mackay, S., Bailey, P.C., Schwinn, K.E., Jameson, P.E., Martin, C.R., and Davies, K.M. (2010). The molecular basis for venation patterning of pigmentation and its effect on pollinator attraction in flowers of Antirrhinum. New Phytol.

Shkolnik, A., Taylor, C.R., Finch, V., and Borut, A. (1980). Why do Bedouins wear black robes in hot deserts? Nature 283, 373–375.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Popul. Biol. 32, 42–49.

Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0. In http://www.repeatmasker.org.

Smithson, A., and Macnair, M.R. (1996). Frequency-dependent selection by pollinators: mechanisms and consequences with regard to behaviour of bumblebees Bombus terrestris (L.) (Hymenoptera: Apidae). J. Evol. Biol. 9, 571–588.

Sokal, R.R. and Michener, C.D. (1958) A Statistical Method for Evaluating Relationships. University of Kansas Science Bulletin, 38, 1409-1448.

Sokal, R.R., and Rohlf, F.J. (1962). the Comparison of Dendrograms By Objective Methods. Taxon 11, 33–40.

Sousa, V., Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. Nat Rev Genet 14, 404–414 (2013). https://doi.org/10.1038/nrg3446

Stephan, W. (2019). Selective sweeps. Genetics 211, 5–13.

Stern, D.L. (2013) The genetic causes of convergent evolution. Nature Reviews Genetics.

Stortenbeker, N., and Bemer, M. (2019). The SAUR gene family: the plant’s toolbox for adaptation of growth and development. J. Exp. Bot. 70, 17–27.

Sutton, D.A. (1988). A revision of the tribe Antirrhineae. (Oxford, UK: Oxford University Press).

Sweigart, A.L., and Willis, J.H. (2012). Molecular evolution and genetics of postzygotic reproductive isolation in plants. F1000 Biol. Rep. 4.

Takahashi, A., Liu, Y.H., and Saitou, N. (2004). Genetic Variation Versus Recombination Rate in a Structured Population of Mice. Mol. Biol. Evol. 21, 404–409.

Tan, Y., Barnbrook, M., Wilson, Y., Molnár, A., Bukys, A., and Hudson, A. (2020). Shared Mutations in a Novel Glutaredoxin Repressor of Multicellular Trichome Fate Underlie Parallel Evolution of Antirrhinum Species. Curr. Biol. 30, 1357-1366.e4.

Tastard, E., Andalo, C., Giurfa, M., Burrus, M., and Thébaud, C. (2008). Flower colour variation across a hybrid zone in Antirrhinum as perceived by bumblebee pollinators. Arthropod. Plant. Interact.

Tastard, E., Ferdy, J.B., Burrus, M., Thébaud, C., and Andalo, C. (2012). Patterns of floral colour neighbourhood and their effects on female reproductive success in an Antirrhinum hybrid zone. J. Evol. Biol.

Tavares, H., Whibley, A., Field, D.L., Bradley, D., Couchman, M., Copsey, L., Elleouet, J., Burrus, M., Andalo, C., Li, M., Li, Q., Xue, Y., Rebocho, A.B., Barton, N.H., and Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. Proc. Natl. Acad. Sci.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S.M., Doebley, J., and Gaut, B.S. (2002). Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). Genetics 162, 1401–1413.

Thakur, V., Wanchana, S., Xu, M., Bruskiewich, R., Quick, W.P., Mosig, A., and Zhu, X.-G. (2011). Characterization of statistical features for plant microRNA prediction. BMC Genomics 12, 108.

Van Belleghem, S.M., Rastas, P., Papanicolaou, A., Martin, S.H., Arias, C.F., Supple, M.A., Hanly, J.J., Mallet, J., Lewis, J.J., Hines, H.M., Ruiz, M., Salazar, C., Linares, M., Moreira, G.R.P., Jiggins, C.D., Counterman, B.A., McMillan, W.O., and Papa, R. (2017). Complex modular architecture around a simple toolkit of wing pattern genes. Nat. Ecol. Evol. 1.

Vargas, P., Carrió, E., Guzmán, B., Amat, E., and Güemes, J. (2009). A geographical pattern of Antirrhinum (Scrophulariaceae) speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. J. Biogeogr. 36, 1297–1312.

Vargas, P., Liberal, I., Ornosa, C., and Gómez, J.M. (2017). Flower specialisation: the occluded corolla of snapdragons (Antirrhinum) exhibits two pollinator niches of large long-tongued bees. Plant Biol. 19, 787–797.

Wang, Y., Wang, X., and Paterson, A.H. (2012). Genome and gene duplications and gene expression divergence: a view from plants. Ann. N. Y. Acad. Sci. 1256, 1–14.

Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., and Benson, G. (2004). Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes. Genome Res.

Webb, D.A. (1971). Taxonomic notes on Antirrhinum L. Bot. J. Linn. Soc. 64: 271–275.

Westerman, E.L., VanKuren, N.W., Massardo, D., Tenger-Trolander, A., Zhang, W., Hill, R.I., Perry, M., Bayala, E., Barr, K., Chamberlain, N., Douglas, T.E., Buerkle, N., Palmer, S.E., and Kronforst, M.R. (2018). Aristaless Controls Butterfly Wing Color Variation Used in Mimicry and Mate Choice. Curr. Biol. 28, 3469-3474.e4.

Whibley, A.C., Langlade, N.B., Andalo, C., Hanna, A.I., Bangham, A., Thébaud, C., and Coen, E. (2006). Evolutionary paths underlying flower color variation in Antirrhinum. Science (80-.).

Wilson, Y., and Hudson, A. (2011). The evolutionary history of Antirrhinum suggests that ancestral phenotype combinations survived repeated hybridizations. Plant J.

Wolf, J.B.W., and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. Nat. Rev. Genet. 18, 87–100.

Wong, T.K.F., Ranjard, L., Lin, Y., and Rodrigo, A.G. (2018). HaploJuice : accurate haplotype assembly from a pool of sequences with known relative concentrations. BMC Bioinformatics 19, 389.

Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution (N. Y). 19, 395.

Xu, P., Billmeier, M., Mohorianu, I.-I., Green, D., Fraser, W., and Dalmay, T. (2015). An improved protocol for small RNA library construction using High Definition adapters. Methods Next Gener. Seq. 2.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: Principles and practice. Nat. Rev. Genet. 13, 303–314.

York, R.A., Patil, C., Abdilleh, K., Johnson, Z. V, Conte, M.A., Genner, M.J., McGrath, P.T., Fraser, H.B., Fernald, R.D., and Streelman, J.T. (2018). Behavior-dependent cis regulation reveals genes and pathways associated with bower building in cichlid fishes. Proc. Natl. Acad. Sci. 115, E11081–E11090.

Zamani, N., Russell, P., Lantz, H., Hoeppner, M.P., Meadows, J.R.S., Vijay, N., Mauceli, E., di Palma, F., Lindblad-Toh, K., Jern, P., and Grabherr, M.G. (2013). Unsupervised genome-wide recognition of local relationship patterns. BMC Genomics 14, 347.

Zhang, Q., Ma, C., Zhang, Y., Gu, Z., Li, W., Duan, X., Wang, S., Hao, L., Wang, Y., Wang, S., and Li, T. (2018). A Single-Nucleotide Polymorphism in the Promoter of a Hairpin RNA Contributes to Alternaria alternata Leaf Spot Resistance in Apple (Malus × domestica). Plant Cell.

Zhang, X., Gonzalez-Carranza, Z.H., Zhang, S., Miao, Y., Liu, C.-J., and Roberts, J.A. (2019). F-Box Proteins in Plants. In Annual Plant Reviews Online, pp. 307–328.

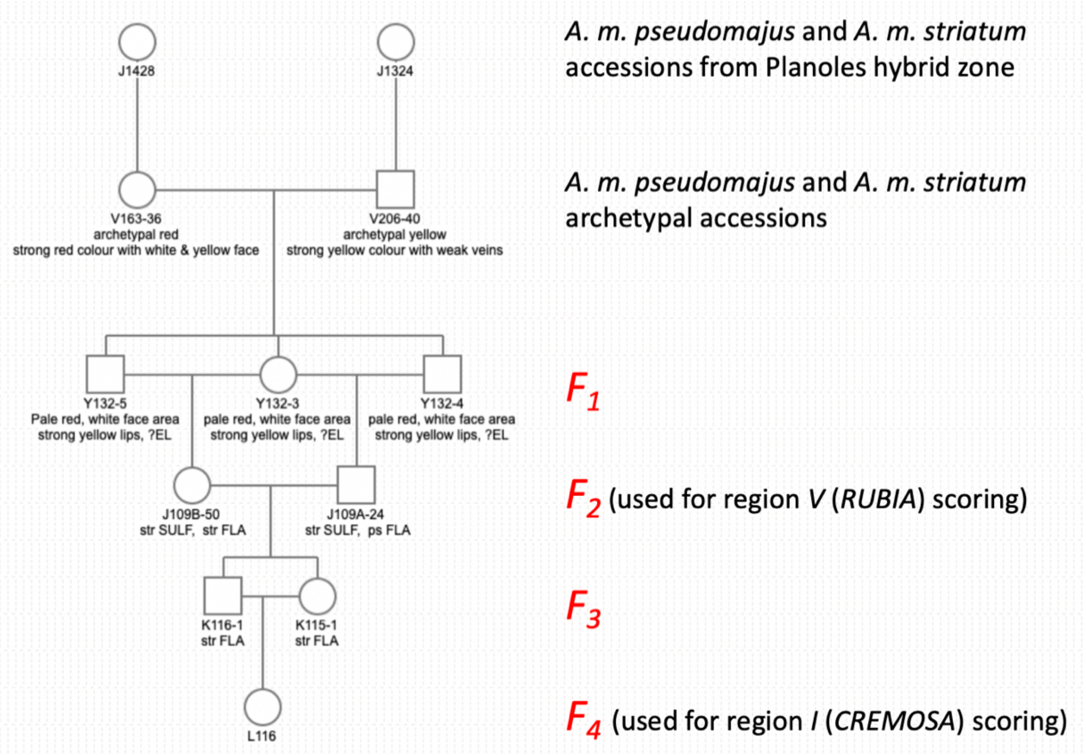
9: Appendices

Appendix 1: GPS coordinates of sampled populations

Population identifier	Number of individuals sampled	Latitude	Longitude
UNA	60	42.763361	1.772739
BED	44	42.869186	1.568953
LU	47	42.968486	2.260464
AXA	32	42.798143	2.2231055
MIJ	44	42.725164	2.039864
MON	50	42.507878	2.122297
PER	43	42.467675	2.8552415
BOU	20	42.643378	2.58705
VIL	50	42.587006	2.367453
ARS	41	42.3895975	2.4876195
THU	47	42.644139	2.721694
BAN	35	42.489458	3.124183
ARL	21	42.4479485	2.6084845
CIN	58	43.311569	1.533579
YP1	50	42.326943	2.052929
YP4	52	42.359921	1.926958
MP4	50	42.322234	2.091375
MP11	50	42.331038	2.170284
W-QUE-A	17	42.11039	1.824566
W-FAI	52	42.16628	1.160356
W-BOX	43	42.1722	1.161639
W-SAL	31	42.22796	1.738268
Néouvielle	32	42.83496	0.159931
Pont Napo	41	42.86	-0.05
Y-VAU	54	43.69867	5.719556
Y-AUR	54	43.36913	5.622545
V-SMP	16	40.41493	-2.76436
V-BUE	57	40.48723	-2.74822

V-CIF	54	40.783	-2.54642
V-PEL	56	41.01442	-2.63701
T-ROZ	15	43.22228	-4.38658
T-HUE	41	42.96974	-4.97417

Appendix 2: Pedigree of F_2 and F_4 families used to investigate *CREMOSA* and *RUBIA* phenotypes


J1428
J1324
A. m. pseudomajus and A. m. striatum accessions from Planoles hybrid zone
V163-36
archetypal red
strong red colour with white & yellow face
V206-40
archetypal yellow
strong yellow colour with weak veins
A. m. pseudomajus and A. m. striatum archetypal accessions
Y132-5
Pale red, white face area
strong yellow lips, ?EL
Y132-3
pale red, white face area
strong yellow lips, ?EL
Y132-4
pale red, white face area
strong yellow lips, ?EL
F1
J109B-50
str SULF, str FLA
J109A-24
str SULF, ps FLA
F2 (used for region V (RUBIA) scoring)
K116-1
str FLA
K115-1
str FLA
F3
L116
F4 (used for region I (CREMOSA) scoring)

Appendix 3: KASP / ALFP oligos used in genotyping flower colour genes

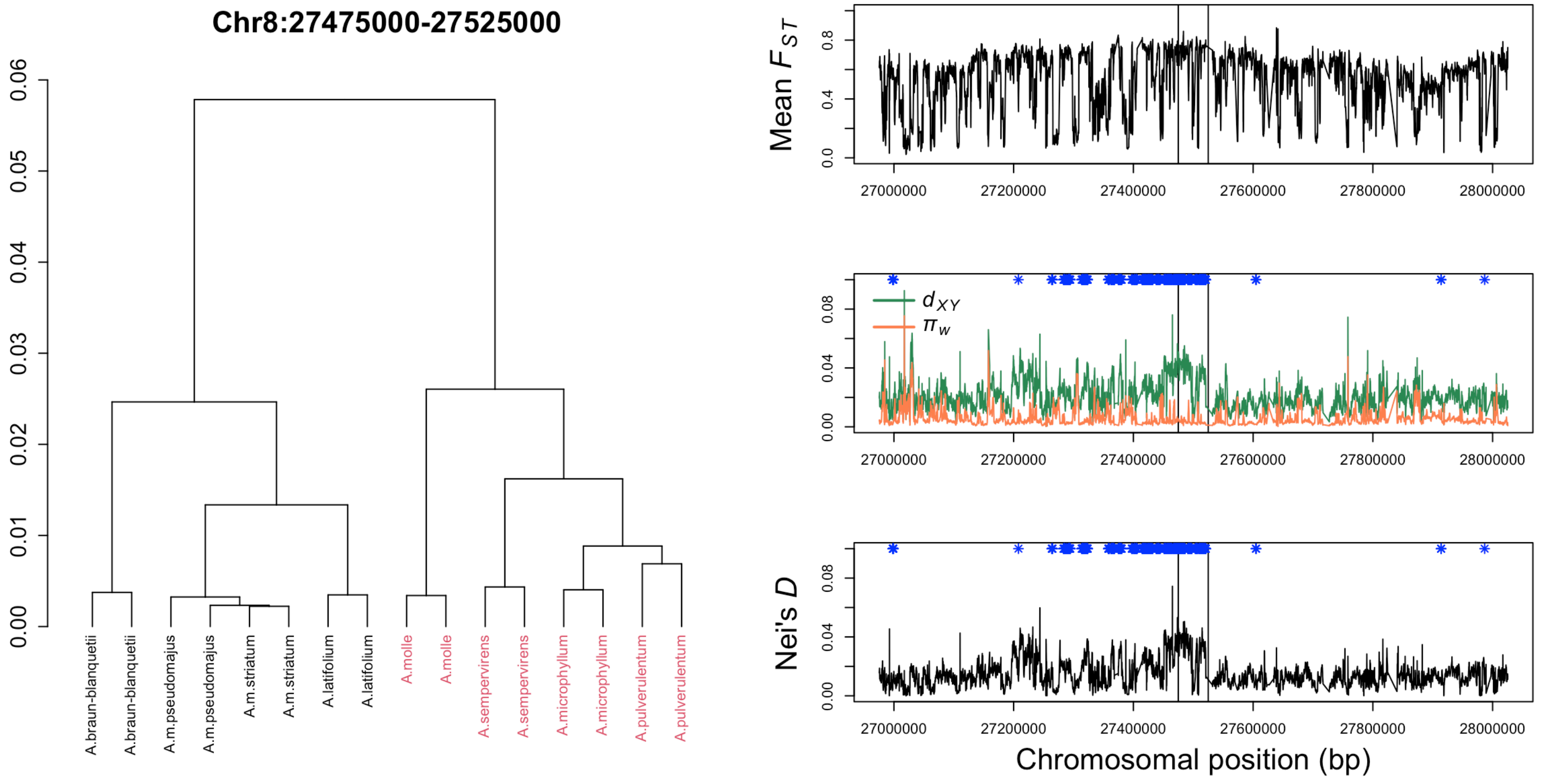
Locus	Marker type	Marker name	Oligo name	Oligo sequence
FLAVIA (upstream)	KASP		#2205	GAAGGTGACCAAGTTCATGCTgattcctcaagcagaaaacg
			#2206	GAAGGTCGGAGTCAACGGATTgattcctcaagcagaaaaca
			#2207	GGAGTGCATCCCTGCCGCG
FLAVIA (downstream)	KASP		do253	GAAGGTGACCAAGTTCATGCTTTCACGTTCTACGAAGGGGTA
			do254	GAAGGTCGGAGTCAACGGATTTTCACGTTCTACGAAGGGGTT
			do255	ctttgcccgttgcttgac
SULF	KASP	Set 65	do514	GAAGGTCGGAGTCAACGGATTGCAAAATCTGCCCTTTTCCAACTT
			do515	GAAGGTGACCAAGTTCATGCTGCAAAATCTGCCCTTTTCCAACTA
			do516	ACTGATGTGAGCGCCGACTGAGC
	KASP	Set 66	do517	GAAGGTCGGAGTCAACGGATTGAATACCACTAAACGAGTGAATGA
			do518	GAAGGTGACCAAGTTCATGCTGAATACCACTAAACGAGTGAATGG
			do519	CTGAATGTCTTCGAAAGGACAGTG
AURINA		Set 61	do502	GAAGGTCGGAGTCAACGGATTTGGAGTCTTAGCGCTCGACACC

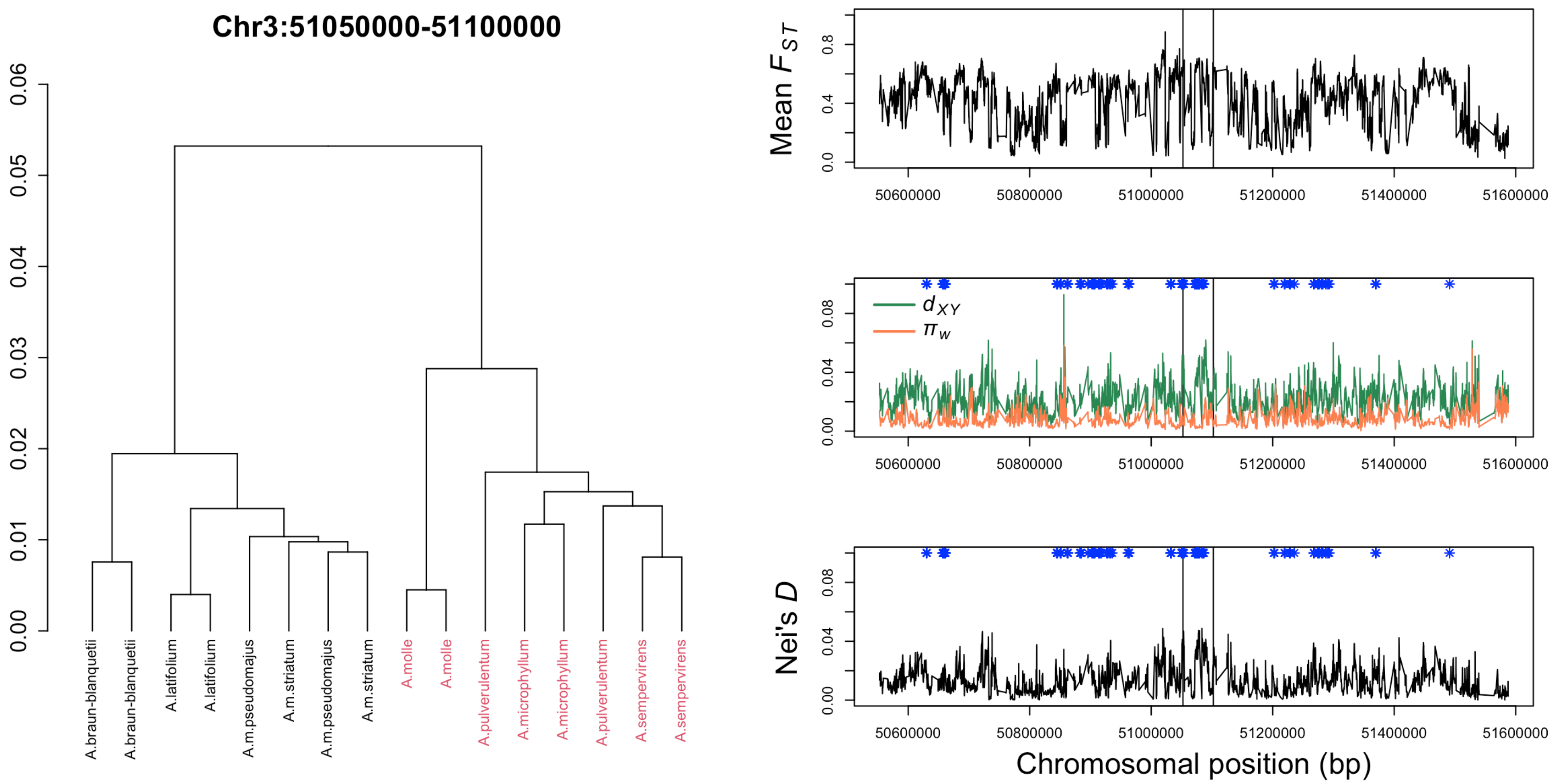
			do503	GAAGGTGACCAAGTTCATGCTTGGAGTCTTAGCGCTCGACACA
			do504	CAATACCACTACTCCTGAAGAGC
CREMOSA		Set 54b	do467	GAAGGTCGGAGTCAACGGATTGTGACTTGGGAGGAAGAATAATC
			d0468	GAAGGTGACCAAGTTCATGCTGTGACTTGGGAGGAAGAATAATA
			do477	TTAAGGGGAAAGTGACTTGATCA
		Do475-476	do475	GAGGCTAGGAAGAAAGGTTTGTCG
			do476	CTAACATTGAGCCAAATATTTGCC
RUBIA		Set 53	do448	GAAGGTCGGAGTCAACGGATTCACACGTGCAGTAATTGAGGCA
			do449	GAAGGTGACCAAGTTCATGCTCACACGTGCAGTAATTGAGGCG
			do50	TTGTTTCAGCTTAAGTTCGGG
ROS1	KASP	ROS1 intron	#1911	GAAGGTGACCAAGTTCATGCTCAACATTGACGTACGGTATTC
			#1912	GAAGGTCGGAGTCAACGGATTCAACATTGACGTACGGTATTT
			#1483	tggcatcaagttccacacagagcag
ELUTA	AFLP		#1615	cattgtcatgactcgttcaaca
			#1616	ttaaactgaaaggcaggcaatc

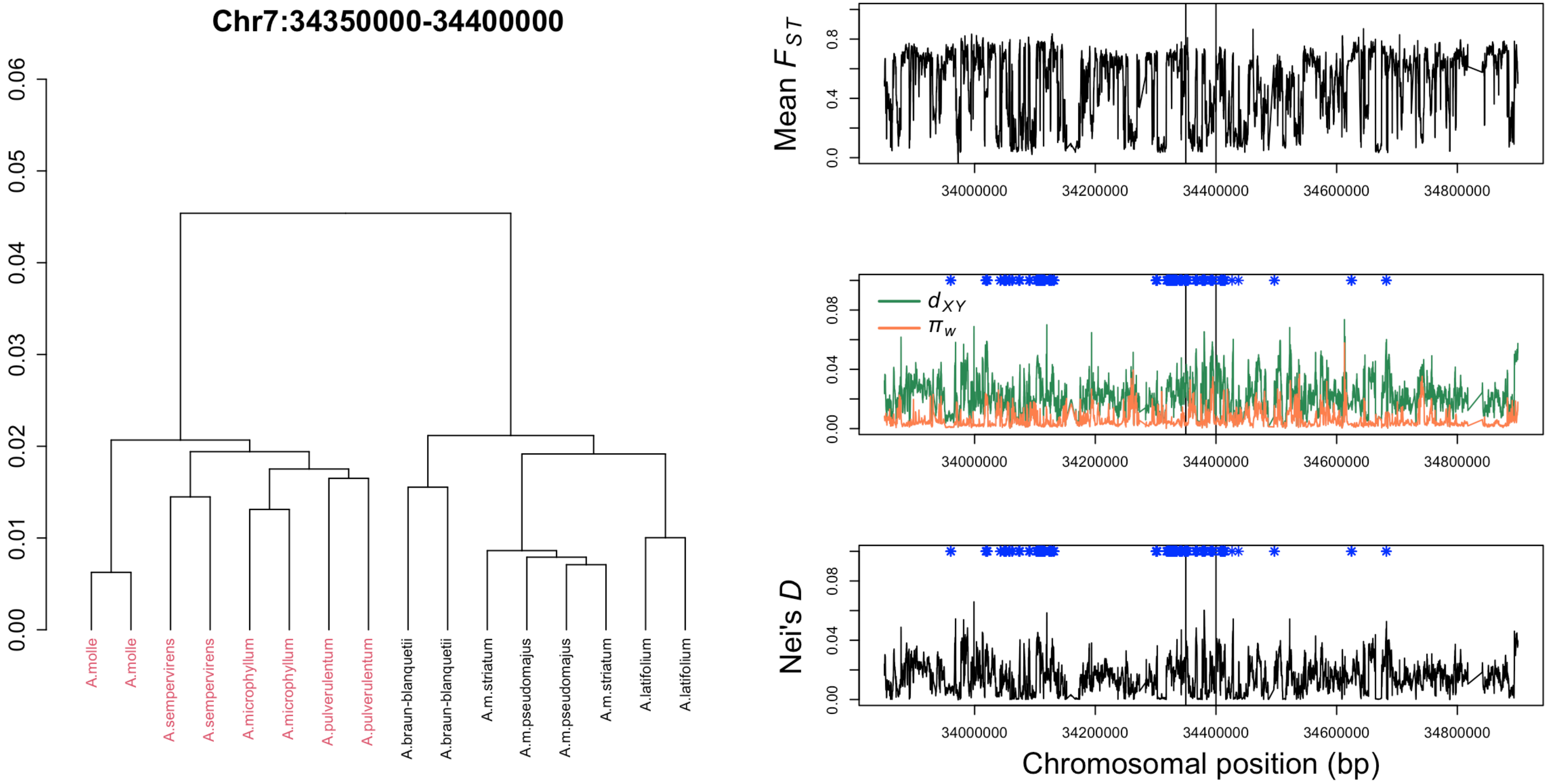
Appendix 4: Depth of genomic coverage across alpine and ruderal species pools

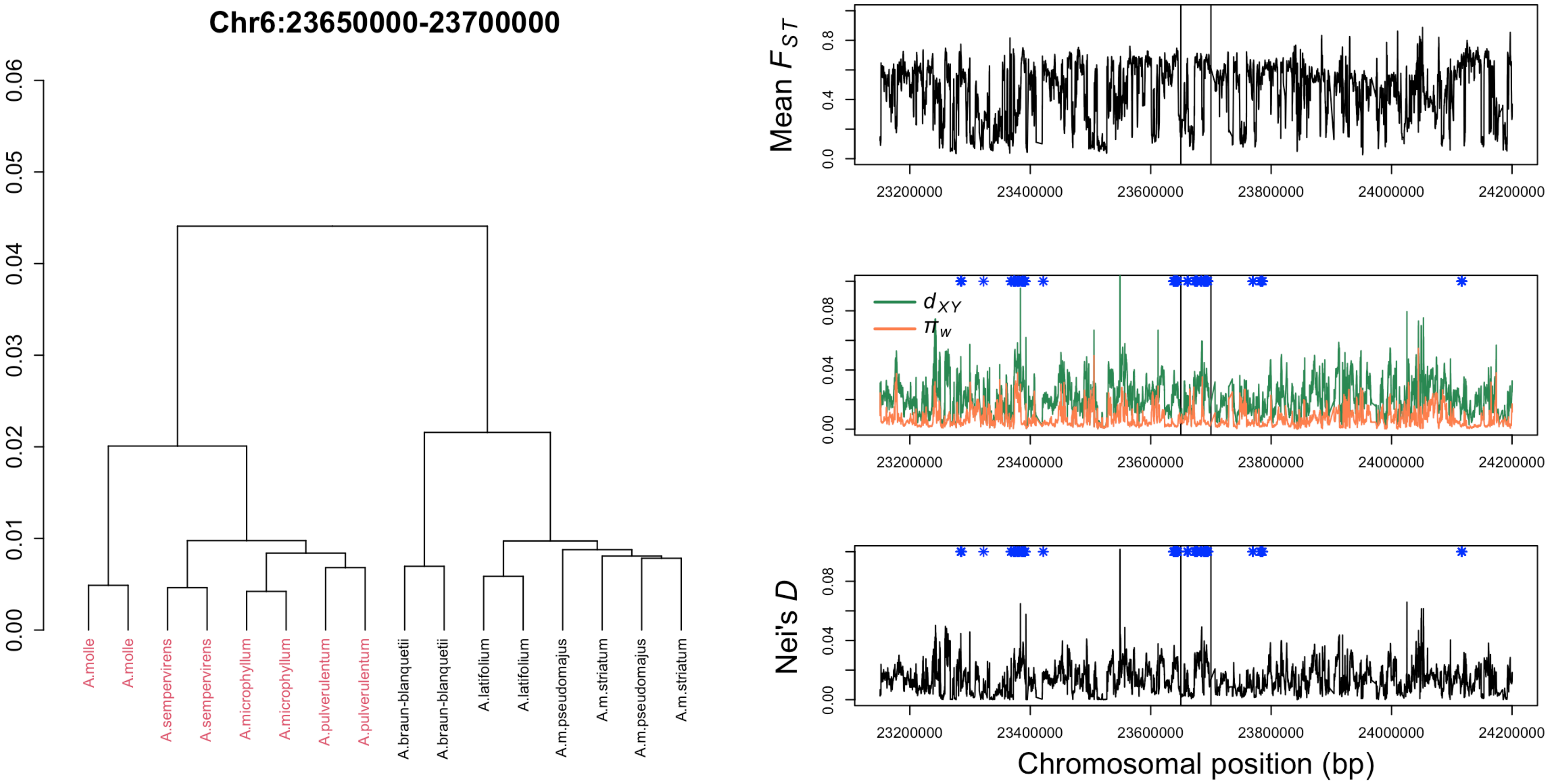
Species	*Location code*	*Growth habit*	*Mean genomic coverage*
A. molle	W-QUE-A	Alpine	66.43
A. molle	W-FAI	Alpine	41.47
A. m. pseudomajus	W-BOX	Ruderal	48.41
A. m. pseudomajus	W-SAL	Ruderal	52.07
A. sempervirens	Néouvielle	Alpine	43.73
A. sempervirens	Pont Napo	Alpine	41.05
A. m. striatum	LU	Ruderal	45.89
A. m. striatum	THU	Ruderal	38.31
A. latifolium	Y-VAU	Ruderal	41.85
A. latifolium	Y-AUR	Ruderal	47.06
A. microphyllum	V-SMP	Alpine	41.93
A. microphyllum	V-BUE	Alpine	36.78
A. pulverulentum	V-CIF	Alpine	42.89
A. pulverulentum	V-PEL	Alpine	38.45
A. braun-blanquetti	T-ROZ	Ruderal	50.96
A. braun-blanquetti	T-HUE	Ruderal	91.04

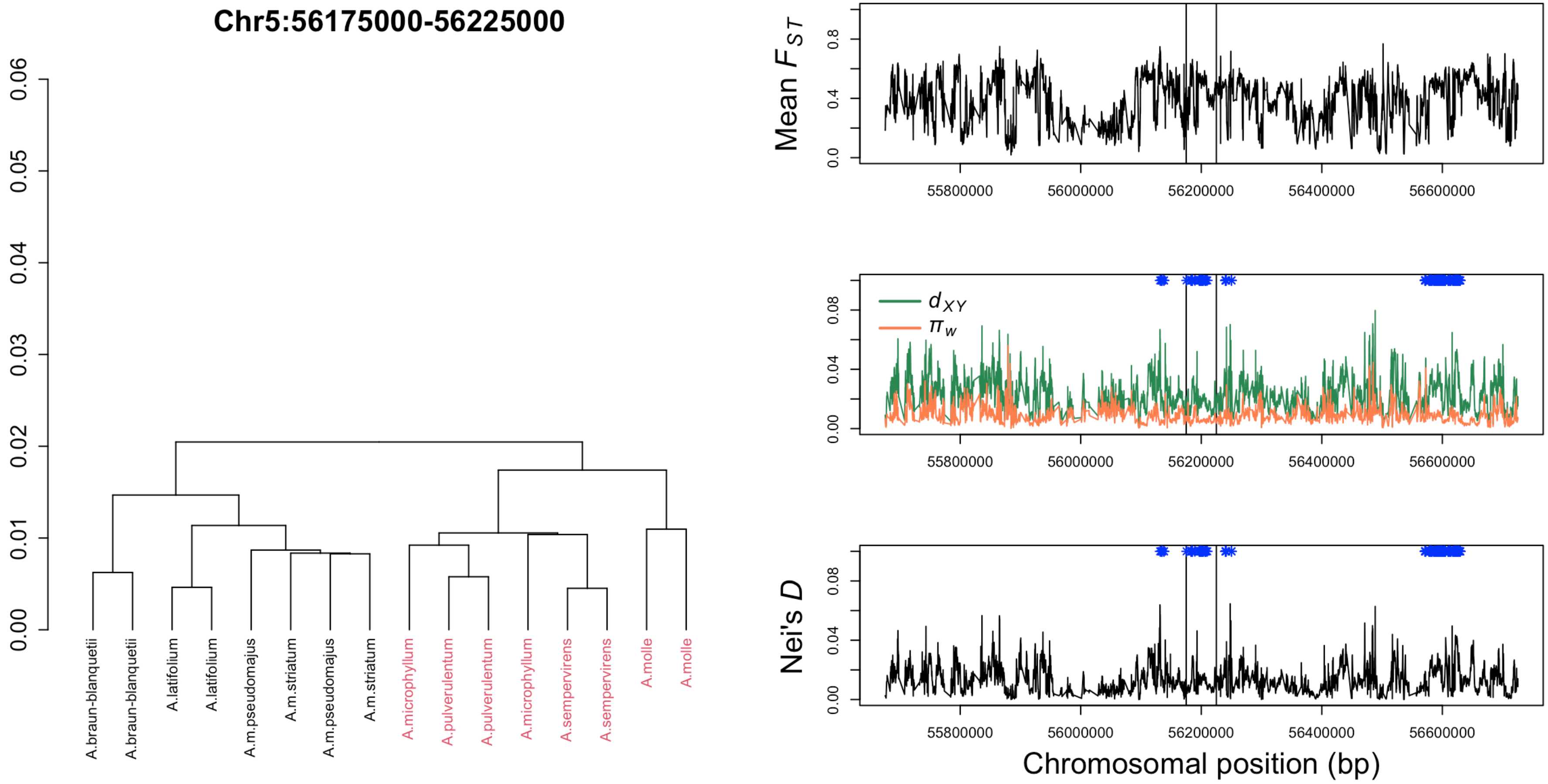
Appendix 5: Plots of F_{ST}, D_{XY}, and π_w across the remaining 12 tested monophyletic regions.

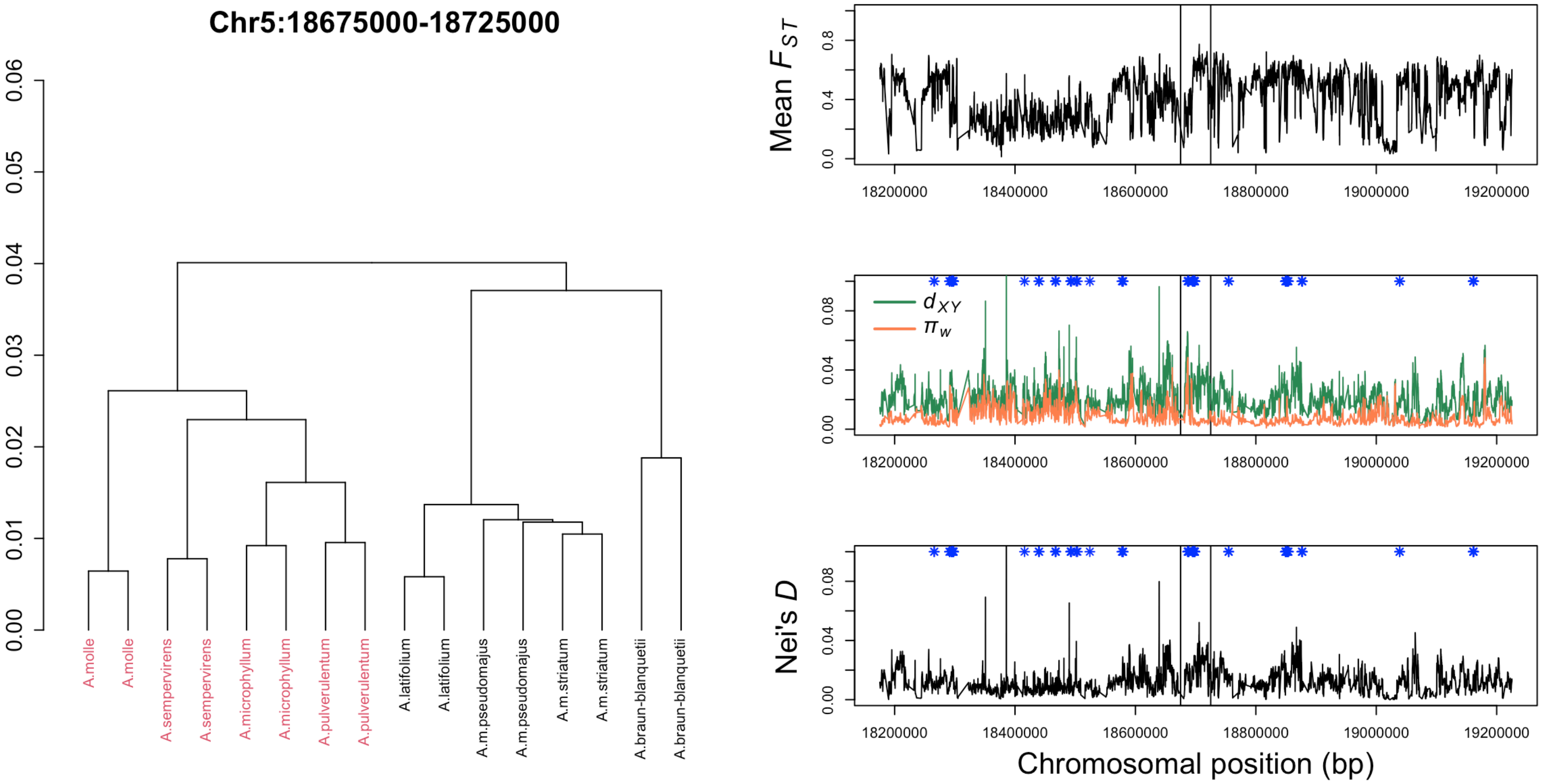

Chr8:27475000-27525000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.braun-blanquetii
A.braun-blanquetii
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
A.m.striatum
A.latifolium
A.latifolium
A.molle
A.molle
A.sempervirens
A.sempervirens
A.microphyllum
A.microphyllum
A.pulverulentum
A.pulverulentum
Mean F_{ST}
0.0
0.4
0.8
27000000
27200000
27400000
27600000
27800000
28000000
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
Chromosomal position (bp)


Chr3:51050000-51100000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.braun-blanquetii
A.braun-blanquetii
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
A.molle
A.molle
A.pulverulentum
A.microphyllum
A.microphyllum
A.pulverulentum
A.sempervirens
A.sempervirens
Mean F_{ST}
0.0
0.4
0.8
50600000
50800000
51000000
51200000
51400000
51600000
d_{XY}
π_w
0.00
0.04
0.08
50600000
50800000
51000000
51200000
51400000
51600000
Nei's D
0.00
0.04
0.08
50600000
50800000
51000000
51200000
51400000
51600000
Chromosomal position (bp)

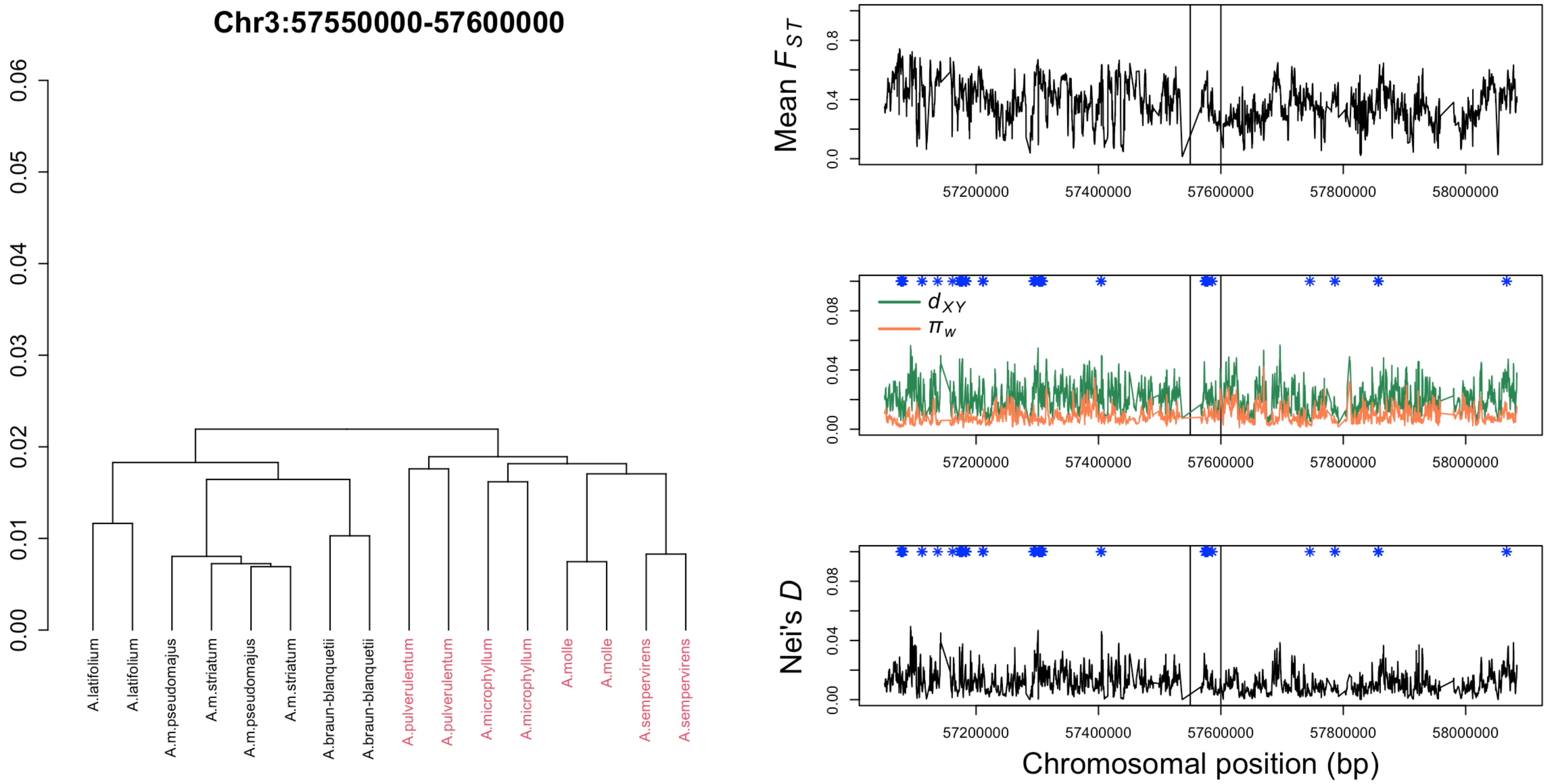

Chr7:34350000-34400000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.sempervirens
A.sempervirens
A.microphyllum
A.microphyllum
A.pulverulentum
A.pulverulentum
A.braun-blanquetii
A.braun-blanquetii
A.m.striatum
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
A.latifolium
A.latifolium
Mean F_{ST}
0.0
0.4
0.8
34000000
34200000
34400000
34600000
34800000
d_{XY}
π_w
0.00
0.04
0.08
34000000
34200000
34400000
34600000
34800000
Nei's D
0.00
0.04
0.08
34000000
34200000
34400000
34600000
34800000
Chromosomal position (bp)

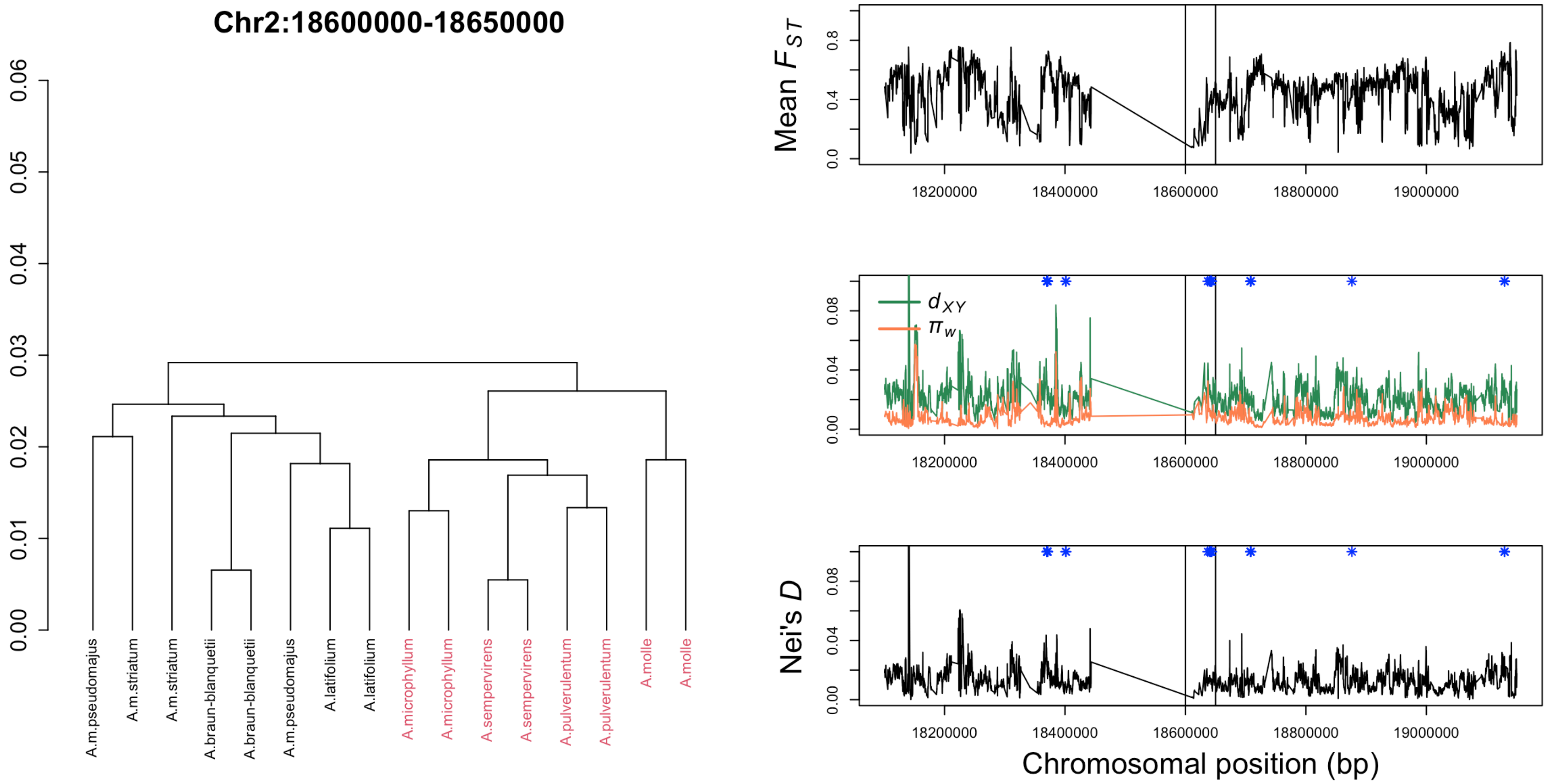

Chr6:23650000-23700000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.sempervirens
A.sempervirens
A.microphyllum
A.microphyllum
A.pulverulentum
A.pulverulentum
A.braun-blanquetii
A.braun-blanquetii
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
0.00
0.04
0.08
23200000
23400000
23600000
23800000
24000000
24200000
Chromosomal position (bp)

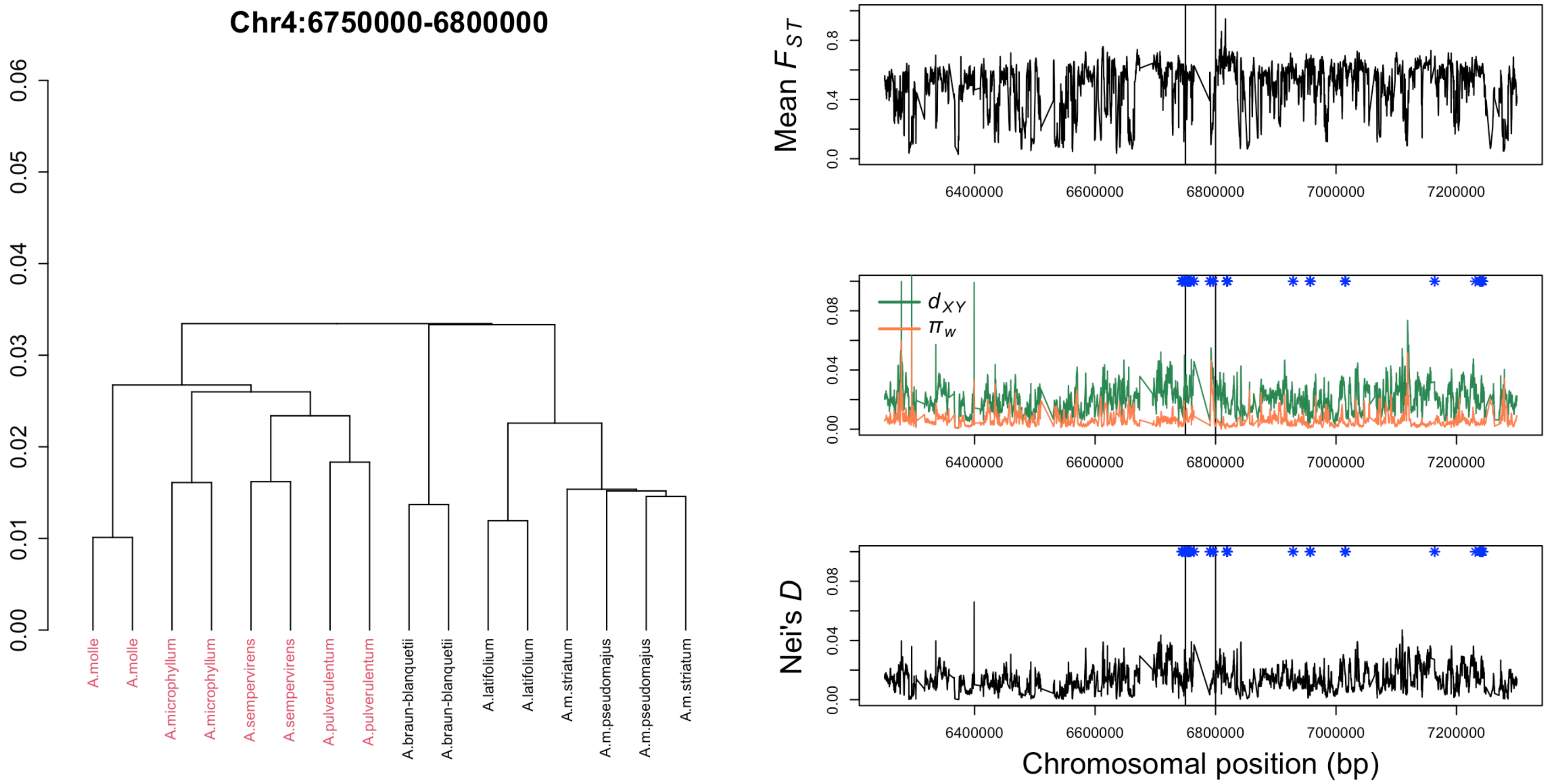

Chr5:56175000-56225000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.braun-blanquetii
A.braun-blanquetii
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
A.microphyllum
A.pulverulentum
A.pulverulentum
A.microphyllum
A.sempervirens
A.sempervirens
A.molle
A.molle
Mean F_{ST}
0.0
0.4
0.8
55800000
56000000
56200000
56400000
56600000
d_{XY}
π_w
0.00
0.04
0.08
55800000
56000000
56200000
56400000
56600000
Nei's D
0.00
0.04
0.08
55800000
56000000
56200000
56400000
56600000
Chromosomal position (bp)

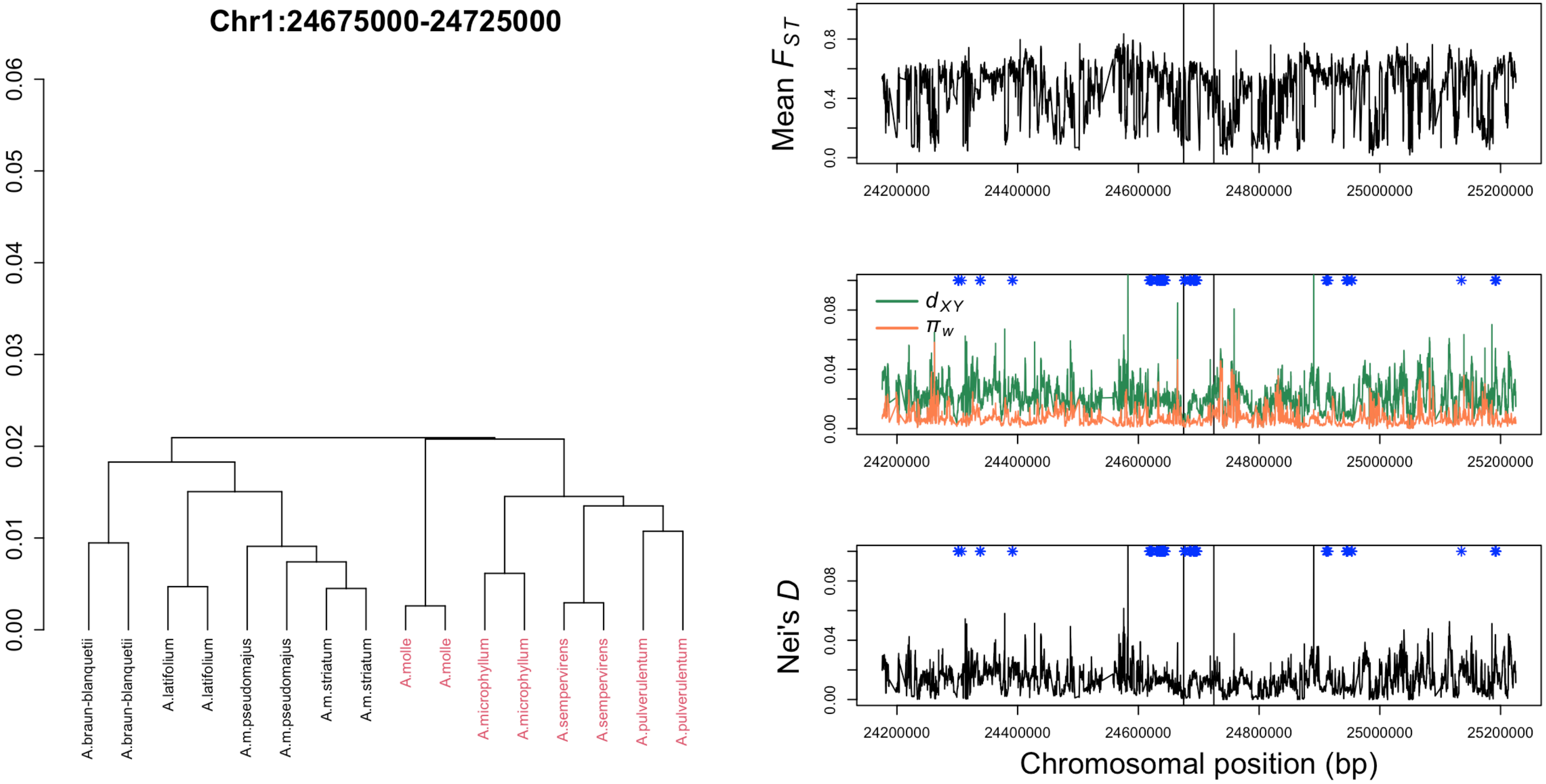


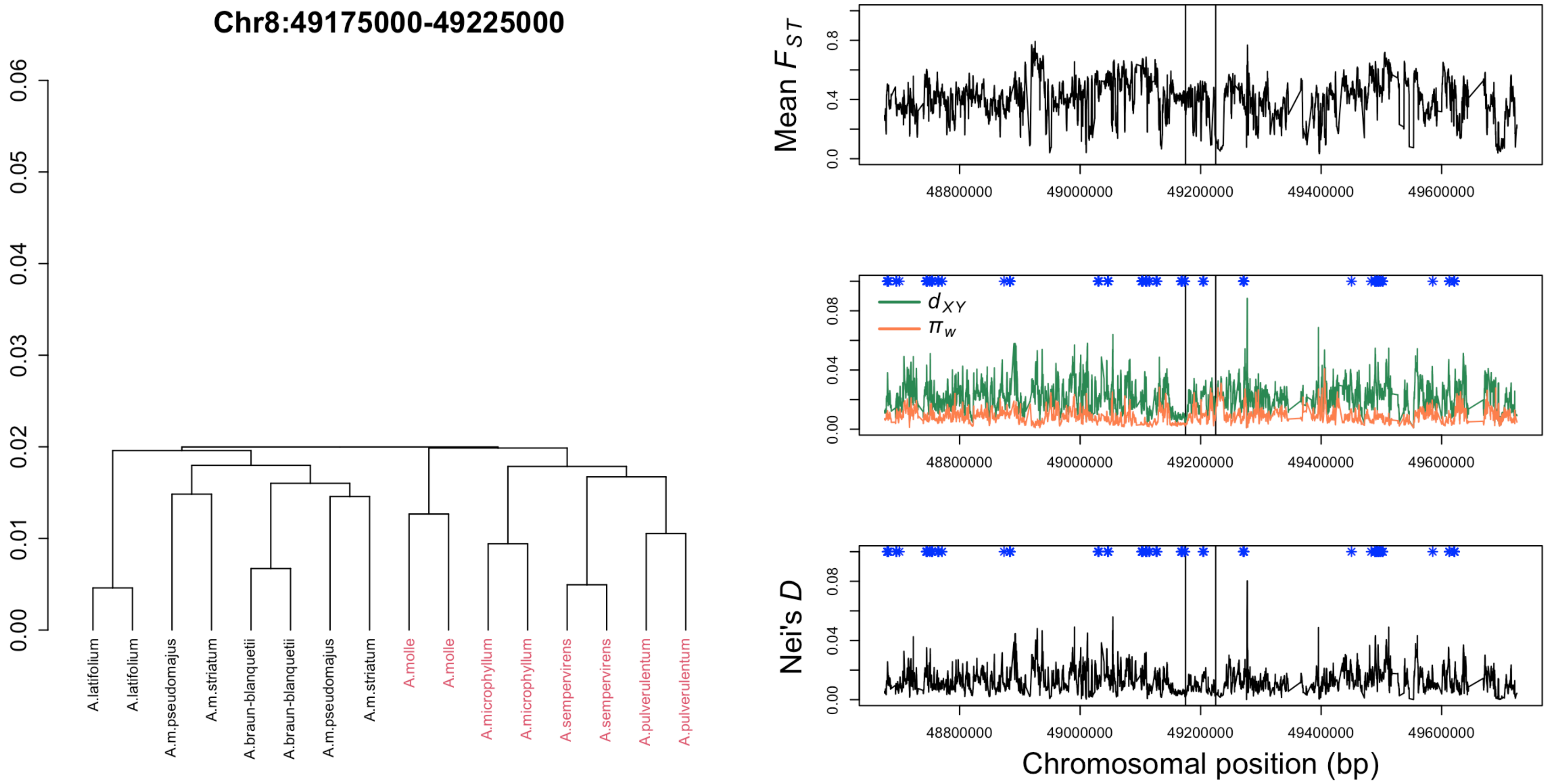
Chr5:18675000-18725000

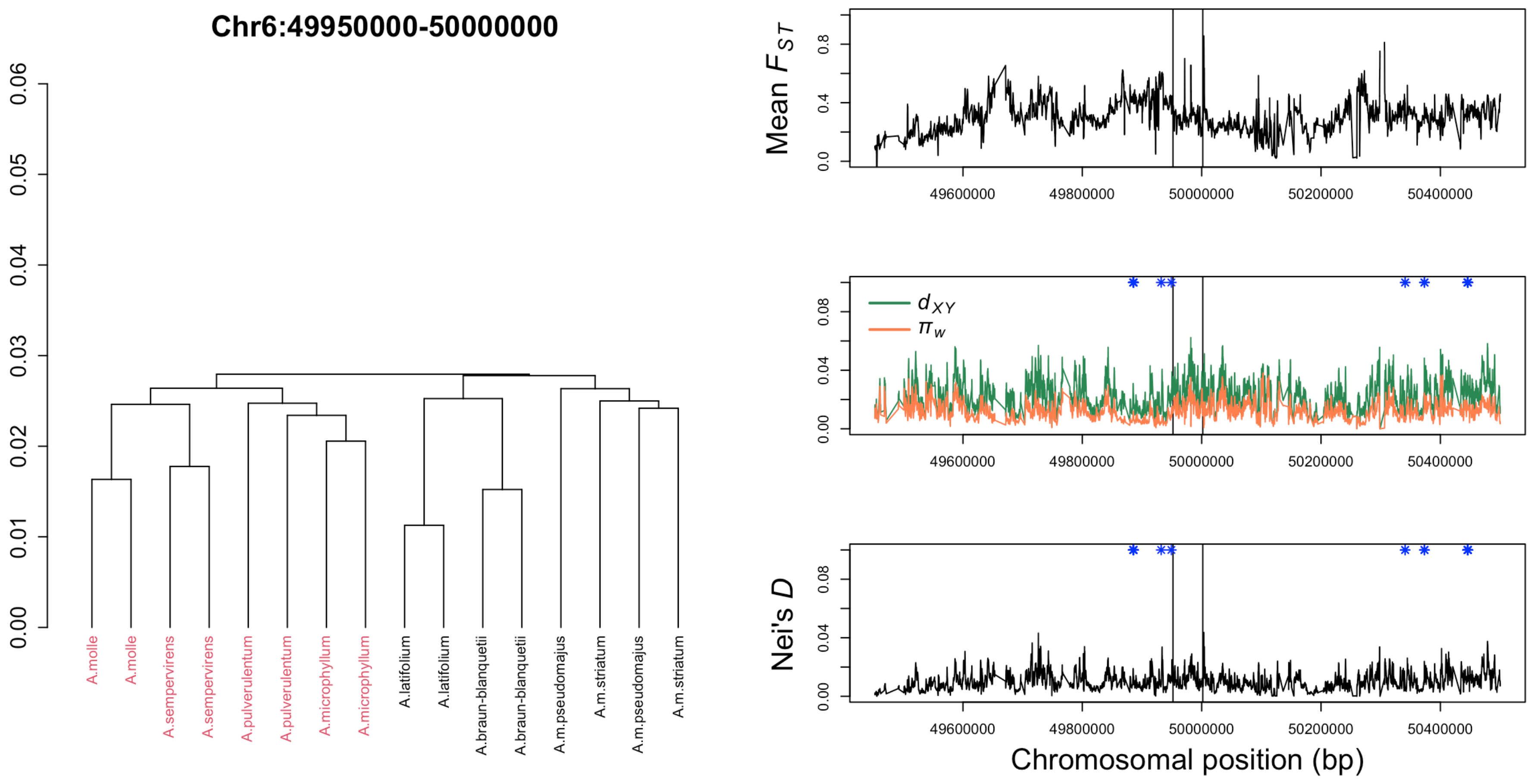

Chr3:57550000-57600000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
A.braun-blanquetii
A.braun-blanquetii
A.pulverulentum
A.pulverulentum
A.microphyllum
A.microphyllum
A.molle
A.molle
A.sempervirens
A.sempervirens
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
0.00
0.04
0.08
57200000
57400000
57600000
57800000
58000000
Chromosomal position (bp)


Chr2:18600000-18650000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.m.pseudomajus
A.m.striatum
A.m.striatum
A.braun-blanquetii
A.braun-blanquetii
A.m.pseudomajus
A.latifolium
A.latifolium
A.microphyllum
A.microphyllum
A.sempervirens
A.sempervirens
A.pulverulentum
A.pulverulentum
A.molle
A.molle
Mean F_{ST}
0.0
0.4
0.8
18200000
18400000
18600000
18800000
19000000
d_{XY}
π_w
0.00
0.04
0.08
18200000
18400000
18600000
18800000
19000000
Nei's D
0.00
0.04
0.08
18200000
18400000
18600000
18800000
19000000
Chromosomal position (bp)


Chr4:6750000-6800000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.microphyllum
A.microphyllum
A.sempervirens
A.sempervirens
A.pulverulentum
A.pulverulentum
A.braun-blanquetii
A.braun-blanquetii
A.latifolium
A.latifolium
A.m.striatum
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
6400000
6600000
6800000
7000000
7200000
Chromosomal position (bp)


Chr1:24675000-24725000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.braun-blanquetii
A.braun-blanquetii
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.pseudomajus
A.m.striatum
A.m.striatum
A.molle
A.molle
A.microphyllum
A.microphyllum
A.sempervirens
A.sempervirens
A.pulverulentum
A.pulverulentum
Mean F_{ST}
0.0
0.4
0.8
24200000
24400000
24600000
24800000
25000000
25200000
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
Chromosomal position (bp)


Chr8:49175000-49225000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.latifolium
A.latifolium
A.m.pseudomajus
A.m.striatum
A.braun-blanquetii
A.braun-blanquetii
A.m.pseudomajus
A.m.striatum
A.molle
A.molle
A.microphyllum
A.microphyllum
A.sempervirens
A.sempervirens
A.pulverulentum
A.pulverulentum
Mean F_{ST}
0.0
0.4
0.8
d_{XY}
π_w
0.00
0.04
0.08
Nei's D
0.00
0.04
0.08
48800000
49000000
49200000
49400000
49600000
Chromosomal position (bp)


Chr6:49950000-50000000
0.00
0.01
0.02
0.03
0.04
0.05
0.06
A.molle
A.molle
A.sempervirens
A.sempervirens
A.pulverulentum
A.pulverulentum
A.microphyllum
A.microphyllum
A.latifolium
A.latifolium
A.braun-blanquetii
A.braun-blanquetii
A.m.pseudomajus
A.m.striatum
A.m.pseudomajus
A.m.striatum
Mean F_{ST}
0.0
0.4
0.8
49600000
49800000
50000000
50200000
50400000
d_{XY}
π_w
0.00
0.04
0.08
49600000
49800000
50000000
50200000
50400000
Nei's D
0.00
0.04
0.08
49600000
49800000
50000000
50200000
50400000
Chromosomal position (bp)