Evolutionary genetics and genomics of flower colour loci in an *Antirrhinum* hybrid zone

Hugo Tavares

PhD Thesis

University of East Anglia

John Innes Centre

Submitted March 2014

Abstract

Unravelling the genetic nature of reproductive isolation is crucial to understanding the maintenance of diversity between populations. In hybrid zones, loci that establish a barrier to gene flow between populations remain divergent, whereas neutral unlinked loci become mixed. In those instances, fit allelic combinations across several loci can be maintained through selection, but this is antagonized by gene flow and recombination. Here, I show that particular allelic combinations in a linked cluster of loci responsible for a flower colour polymorphism between two *A. majus* subspecies are maintained despite recombination in a hybrid zone. I reveal that a known locus that controls the magenta flower colour of the subspecies, *ROSEA* (*ROS*), is highly divergent between them, compared with most of the genome. The divergence region extends downstream of *ROS*, likely due to selection on another linked, but unidentified, locus that also controls flower colour, *ELUTA* (*EL*). Fine-mapping experiments identified an interval containing *EL* and regions within *ROS* that control different components of the magenta phenotype. Transcriptome analysis from flower buds suggests that *MYB*-like transcription factors within the mapped intervals control this trait. *ROS* and *EL* interact epistatically, meaning that the phenotype of an individual depends on the particular allelic combination it has for these loci. In the hybrid zone, markers in *ROS* and *EL* are in high linkage disequilibrium, but ~5% of recombinant haplotypes were found in the population. Recombinant haplotypes modify the phenotype of the flowers in relation to the parental subspecies, and therefore may be selected against. The data suggest that allelic combinations in *ROS-EL* are maintained by selection, despite gene flow and recombination between the two subspecies. This work reveals the consequences of selection, gene flow and recombination in shaping the patterns of genomic divergence in linked clusters of loci that establish an isolating barrier between populations.

Contents

List of figures

List of tables

List of accompanying material

Attached is a CD, containing:

- A Microsoft Excel file named “mapping_recombinants.xlsx”, which contains the genotypes for the *ROS-EL* recombinants used in mapping experiments, detailed in Chapter 4.

- A Microsoft Excel file named “hz_haplotypes.xlsx”, which contains the genotypes for *ROS-EL* recombinant haplotypes collected from the hybrid zone, detailed in Chapter 6.

- A Microsoft Word file named “figures.docx”, which contains all the figures in this thesis in high resolution.

Author contributions to results chapters

Chapter 3

Xana Rebocho and Desmond Bradley collected leaf material in the field, used for whole-genome sequencing. Desmond Bradley did the pooled DNA extraction of those samples. Annabel Whibley did the assembly of a BAC containing *ROS1*. All sequencing libraries and runs were provided as service by TGAC. I did all of the remaining bioinformatics and data analysis presented in the results section.

Chapter 4

Lucy Copsey produced an F2 between *A. m. pseudomajus* x *A. m. striatum* analysed in section 4.2.1. Lucy Copsey and Annabel Whibley screened several *ROS-EL* recombinants in 2009 (section 4.2.2). Lucy Copsey provided support with all of the glasshouse work. I did all other recombinant screens, genetic crosses, molecular genotyping and data analysis presented in the results section of this chapter.

Chapter 5

Sequencing libraries and runs for RNAseq were provided as a service by TGAC. I did the harvesting of tissue, RNA extraction, bioinformatics and data analysis detailed in the results section.

Chapter 6

The leaf material used from the hybrid zone was collected in the field between the years of 2009-2012 by myself and several colleagues and volunteers. The DNA extraction of those samples was provided as a service by Richard Goram. Joanne Ellequet genotyped 288 of these samples with *ROS1* and *EL-MYB* markers and did an initial screen for recombinants (the remaining 2107 samples were genotyped by me). Desmond Bradley helped in the collection of capsules from the hybrid zone. Lucy Copsey did all of the test-crosses presented in sections 6.2.4 and 6.2.5. I did all the remaining molecular genotyping and all of the genetic and phenotypic analyses presented in the results section.

Personal acknowledgements

Many thanks to:

My supervisor, Enrico Coen, for his continuous support and ideas, which guided this thesis project.

Desmond Bradley, for too many things, but most importantly for his endless joy and uplifting support, which brought me to the completion of this thesis.

Lucy Copsey, who constantly helped with the logistics of caring for the thousands of *Antirrhinum* plants grown for this work.

Matt Couchman, for helping with bioinformatics and managing the humongous hybrid zone database.

Joanne Elleouet, who started the work on finding recombinants in the hybrid zone and did a lot of the preliminary work on analysing genomic datasets from the populations.

Louis Boell, who joined me in the continuous struggle to make sense of the genomic datasets and is always available for (fun and often inconclusive) discussions.

Nick Barton, who clarified many aspects of my data and helped with their analysis.

David Field, who helped with developing SNP markers for the hybrid zone and much of the field work.

Annabel Whibley, who gave many sapient advices, engaged in interesting discussions about this work and helped with bioinformatics.

Xana Rebocho, who helped me a lot in the early stages of this project, and always kept an enthusiasm that pushed me forward in my PhD.

Yongbiao Xue, who provided early access to the *Antirrhinum* genome reference, bringing this project to the "genomic era".

Katie Abley who read all the early (sometimes disastrous) versions of this thesis' chapters and patiently guided me in a better direction.

The "Tolouse Team", Monique Burrus, Christophe Thébaud, Christophe Andalo for always being available to help with field work and for keeping ecology alive in the hybrid zone project.

My secondary supervisor, Caroline Dean, for many constructive advices on this work.

Many students, friends and colleagues who helped over the years with the field work in the hybrid zone.

1 General introduction

One of the main goals in evolutionary biology is to understand the origin and maintenance of trait diversity in natural populations. A population changes over time by the accumulation of several genetic changes, which might become fixed due to the effect of natural selection (which favours a genetic variant over another) or simply by the random process of genetic drift (where a genetic variant becomes fixed by chance). As different populations with a common ancestor change over time, they will accumulate distinct genetic changes and increasingly diverge from each other. If enough differences accumulate between such diverging populations, they may become established as distinct species that no longer exchange genetic material with each other (i.e. there is complete reproductive isolation).

During the process of divergence, populations might go through many demographic changes: populations might expand or contract, individuals may colonize new habitats, migrate between populations, etc. Therefore, populations do not always occur as geographically separated units. Instead, the history of divergence between lineages may be punctuated by events of migration and hybridization, and these events are likely to play a considerable role in the speciation process (Soltis & Soltis 2009; Abbott *et al.* 2013). Therefore, two populations may have diverged with regard to some character (or characters), but still be inter-fertile if they meet with each other (these are in a taxonomic “grey area” and may be considered as different species, subspecies, varieties, races, ecotypes, etc.).

If divergent populations are still inter-fertile, then an isolating barrier (or barriers) that reduces gene flow between them must exist if they are to remain as separate entities. If such barriers do not exist, then the two populations' gene pools will mix and any fit allelic combinations that have been established in the two separate populations will be broken down by recombination. In other words, gene flow and recombination have a homogenizing effect that counteracts the divergence process that established any differences to start with (Barton & Hewitt 1989; Abbott *et al.* 2013). How then, can diversity be maintained in the face of gene flow?

Several kinds of isolating barriers between populations may evolve, which reduce the extent of gene flow during periods of hybridization (Bolnick & Fitzpatrick 2007; Lowry *et al.* 2008). The nature of these barriers can be varied, but they are of a genetic nature and involve some form of selection. Studying such reproductive barriers can therefore elucidate the genetic basis of divergence and, ultimately, reproductive isolation. The identification of the actual genes behind these isolating barriers (sometimes referred to as "speciation genes") might uncover several aspects of this divergence. At the molecular level, it reveals whether mutations occur in the coding sequence of genes or on their regulatory elements, and whether those genes encode enzymes in the relevant pathways or regulatory transcription factors that regulate their expression (Carroll 2005; Rieseberg & Blackman 2010). At the ecological level, it clarifies if "speciation genes" are responsible for adaptation to local environments or if they establish intrinsic genetic incompatibilities between populations, independently from the environment (Orr *et al.* 2004; Rieseberg & Blackman 2010). And finally, knowledge of the genes might pinpoint the evolutionary forces that shape this divergence, namely if divergence is established primarily due to stochastic forces or if it involves positive selection (Schluter 2009).

A reproductive barrier restricts gene flow for alleles in loci that control the trait under selection; however, alleles in other loci in the genome may freely flow between the populations if they hybridize. In this sense, reproductive isolation between populations might not be complete (i.e. affecting the whole genome), but differences between populations can still be maintained because gene flow is reduced for alleles in those loci that establish the reproductive barrier (Orr 2001; Lexer & Widmer 2008). Because reproductive isolation is not established by the entire genome, it is important to identify the individual loci that cause a reproductive barrier and clarify how their allele frequencies change in face of hybridization and gene flow. Further, in the context of a genome, it becomes relevant to consider how the counteracting effects of selection and recombination respectively increase or decrease the divergence of the barrier loci compared to other loci in the genome.

This work considers these questions by detailing the genetic basis of a flower colour polymorphism between two hybridizing *Antirrhinum majus* subspecies. I analysed a cluster of linked loci that, together, significantly contribute to the difference in flower colour between the two subspecies. By using a combination of genetic, genomic and population genetics analyses, I address how natural selection maintains this trait difference, despite gene flow and recombination in a hybrid zone between the two subspecies.

1.1 Genetic incompatibilities as barriers to gene flow

There are many kinds of barriers that can impede or reduce gene flow between inter-fertile populations. An obvious case is geographic separation, which reduces the number of migrants between populations. This can occur, for example, if two populations are distant from each other (significantly further than the average dispersal distance of the individuals) or if some insurmountable landscape features (e.g. mountains, rivers, oceans) separate them. This kind of barrier is not of a genetic nature and does not necessarily involve any role for selection, as genetic differences between populations may accumulate simply by random processes (drift) (Ridley 2004). The focus of this work, however, is those cases where inter-fertile populations occur in sympatry or parapatry and thus might hybridize. In those cases, other kinds of barriers to gene flow are involved that require some form of selection to establish a reproductive incompatibility between the individuals from the two populations (Bolnick & Fitzpatrick 2007; Lowry *et al.* 2008). These barriers may involve incompatibilities caused by specific ecological adaptations, incompatibilities in the mating system, or incompatibilities caused by unfavourable genetic interactions.

Ecological adaptation to local environments reduces gene flow between two populations because individuals have a loss in fitness when occurring in a non-native habitat. For example, *Mimulus cardinalis* and *M. lewisii* species are inter-fertile, but there is a correlation between each species' habitat range and their fitness (Angert & Schemske 2005). While *M. cadrinalis* thrives in regions of lower altitude, *M. lewisii* inhabits regions of higher altitude. Swapping members of each population to the habitat of the other, results in a significant decrease in fitness for the transplanted individuals (fitness was measured as each individual's survival, growth and number of flowers produced). A similar example occurs with the eco-geographic separation of two *Gilia capitata* subspecies that inhabit coastal and inland regions (Nagy & Rice 1997). Again, each subspecies is better adapted to its own environment, with individuals showing reduced fitness when transplanted to a non-native environment (in this case, fitness was measured as the number of seedlings that emerged in each environment, number of flowering individuals and number of flowers per individual). In both cases, local adaptation results in fitness costs for non-native individuals, thus contributing to maintain a barrier between populations, even if the hybrids are viable and fertile.

Incompatibilities in the mating system result in another kind of reproductive barrier between populations. In plants, this is often associated with differences in floral traits, such as flower colour (detailed later) or other mating system disparities (Lowry *et al.* 2008). For example, two inter-fertile *Iris* species (*I. fulva* and *I. brevicaulis*) that occur in the same geographic regions (sympatry) are mostly isolated due to an asynchronous flowering time (Cruzan & Arnold 1994; Martin *et al.* 2007). The genetic basis of this difference was mapped to several quantitative trait loci (QTL). Reciprocal backcrosses between the two species revealed that QTL from the late-flowering species introgressed into the early-flowering species caused flowering to occur later, while the reciprocal cross showed an opposite effect (Martin *et al.* 2007). In this example, the shift in the mating time results in a significant reduction of gene flow between the two *Iris* species, even if they occur in sympatry.

Unfavourable genetic interactions can also establish a barrier to gene flow. In this case, intrinsic incompatibilities between genotypes of individuals from different populations reduce the fitness of hybrids (which may have reduced fertility, complete sterility or premature death). The evolution of these hybrid incompatibilities can be explained by the Bateson-Dobzhansky-Muller (BDM) model (Orr 1996; Ridley 2004). The model states that hybrid incompatibilities arise from negative interactions between more than one locus in the genome (i.e. negative epistasis). Essentially, different mutations can accumulate in separate populations, as long as they are harmless (no effect on fitness) or advantageous (positively selected) in the context of each population's genetic background. However, different populations might accumulate mutations that, when brought together in a hybrid genome, result in a fitness loss. Although adaptation to divergent habitats (mentioned above) might lead to the establishment of these incompatibility alleles by selection, these might also occur as a consequence of internal genomic conflicts that evolve differently in separate lineages (e.g. silencing of transposable elements or heterochromatin stability) (Michalak 2009; Presgraves 2010).

An example of a BDM incompatibility, associated with the evolution of immune response in plants, is the occurrence of necrosis observed in F1 hybrids between *Arabidopsis thaliana* individuals from different natural origins (Bomblies *et al.* 2007). Two unlinked loci responsible for this phenotype were mapped, both having numerous sequence differences between incompatible individuals. A transformation assay revealed that an allele from a single locus (the pathogen-response gene *NB-LRR*), was enough to trigger a necrotic phenotype in a non-compatible genetic background. This shows how genetic differences between populations (even if they are of the same species) may lead to genetic incompatibilities in a hybrid genomic environment where divergent alleles are meeting for the first time.

These examples illustrate how partial reproductive isolation between populations occurs through the establishment of barriers that have a genetic basis. Often, these barriers are not due to the effect of a single locus, but rather multiple loci that interact with each other to produce viable and fit phenotypes within a population. In this context, the sets of loci that establish the reproductive barrier can be seen as a cohesive co-adapted unit that is maintained intact by selection (Wallace 1991).

1.2 Occurrence of co-adapted loci in natural populations

The term co-adaptation was introduced by Dobzhansky (1950) to refer to the fitness effects of chromosomal inversions in different *Drosophila pseudoobscura* populations. By making crosses between flies from two different populations, he observed that the fitness of the individuals significantly decreased if they carried an inversion allele from each population. He suggested that the inversion polymorphisms contained multiple loci that were “mutually adjusted” or “coadapted” within each population, but deleterious interactions occurred if the inversion polymorphisms from different populations occurred together in the same genome (Dobzhansky 1950). More generally, co-adaptation is defined as the occurrence of allelic combinations in a set of loci that generate fit phenotypes within the context of a particular population. The nature of this co-adaptation can be revealed in hybrids between divergent populations, where new allelic combinations result in a loss of fitness compared with the parental genotypes.

Heliconius butterflies provide an example of co-adapted loci that control the extensive variation in wing colour between hybridizing races of several species (Mallet & Joron 1999). There is strong evidence that colour patterns are correlated with assortative mating in this species complex, such that males are more likely to recognize females with a similar wing pattern to their own (Merrill *et al.* 2011, 2014). Although this behavioural barrier is not absolute, it suggests that individuals with recombinant phenotypes might be intrinsically less fit because they might not be recognized for mating by the more common morphs. If this is the case, there is fitness epistasis between loci controlling mate preference and wing coloration (Jiggins *et al.* 2004). An additional mode of selection operates on butterfly wing patterns, which is unrelated to mate recognition. Generally, within a given geographic location the wing pattern is shared across several species, a strategy of Müllerian mimicry (a warning signal for predators that is shared between distasteful species). If an individual has a wing pattern that differs from the common pattern in a particular location, it will incur a fitness loss, since it will not be recognized as unpalatable by the local community of predators (positive frequency-dependent selection). Therefore, although hybrid zones can form between races with different wing patterns, the parental phenotypes are maintained due to selection against hybrid phenotypes, either due to assortative mating or frequency-dependent selection by predators (Mallet *et al.* 1990; Baxter *et al.* 2010; Counterman *et al.* 2010).

The *Heliconius* example reveals an important feature of the genetic nature of isolation: that the differentiation between populations is not due to an incompatibility established by the whole genome, but rather by a set of loci that interact to establish that incompatibility. These loci are expected to remain highly divergent between populations because their introgression to a foreign genetic background is maladaptive. But what happens to the rest of the genome? If hybrids between populations are still viable and fertile, they may eventually cross with other individuals. Therefore, loci that do not contribute to establish a reproductive barrier, and thus are not under a strong selective pressure, might be exchanged between populations (Barton & Hewitt 1989; Abbott *et al.* 2013). This disparity between the rate of gene flow of alleles in selected and non-selected loci is expected to result in a heterogeneous pattern of divergence across the genome (Feder *et al.* 2012; Nosil & Feder 2012).

1.3 Patterns of divergence across genomes

In recent years, access to genome-wide polymorphisms from natural populations has revealed how selection, gene flow and recombination shape the patterns of divergence in the genome (Feder & Nosil 2010; Nosil & Feder 2012). In cases where gene-flow occurs, the divergence levels are mostly low across the genome, reflecting the history of genetic-exchange between the populations. In contrast, loci under selection (those that establish a reproductive barrier) are highly differentiated, since they do not introgress from one population to the other. This results in a heterogeneous pattern of divergence across the genome that has been metaphorically described as containing "islands" of high divergence in a "sea-level" of lower divergence (Figure 1.1). The "islands" are the result of selection acting on certain loci that are not exchanged between populations (thus divergence is high), whereas the "sea" reflects shared ancestral variation and/or a long history of hybridization and gene exchange (thus divergence is lower).

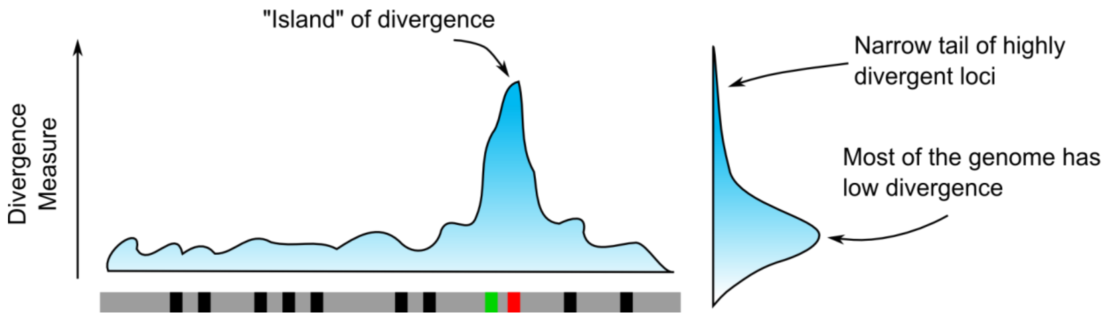

"Island" of divergence
Divergence Measure
Narrow tail of highly divergent loci
Most of the genome has low divergence

Figure 1.1 – Illustration of the concept of a genomic "island" of divergence.
Typically, if two populations are only differentiated in a few loci, the distribution of divergence across the genome will be skewed (shown on the right), that is, most of the genome will have relatively lower divergence than those regions. This variation can be visualized by plotting a nucleotide divergence measure across the genome (genome represented as a grey bar). A locus that is under divergent selection (red box) will result in locally elevated divergence levels. Due to linkage, the divergence extends around the selected locus, such that nearby neutral loci (green box) will also be included within the "island". Elsewhere in the genome, the levels of divergence fluctuate around the average (neutral loci shown as black boxes).

In the aforementioned case of divergent wing-pattern phenotypes of *Heliconius* butterflies, the genomic heterogeneity of divergence is well studied (Baxter *et al.*

2010; Joron *et al.* 2011; Nadeau *et al.* 2012, 2013). For example, in races of *H. melpomene* that are known to hybridize with each other, the divergence is generally low across the genome, but not in the regions that contain loci known to control wing colour patterning. Another recently studied case occurs between two flycatcher bird species, *Ficedula albicollis* and *F. hypoleuca* (Ellegren *et al.* 2012). The history of these species has been marked by alternating periods of allopatry and sympatry, the latter resulting in the formation of hybrid zones that can still be found today (Saetre *et al.* 1999). The current and past events of gene flow might have homogenized the genome divergence, except for a few regions that form islands of high divergence (Ellegren *et al.* 2012). A final example is the case of population divergence with gene flow in the fly *Rhagoletis pomonella* (Feder *et al.* 2003; Michel *et al.* 2010). The larvae of this fly feed on the flesh of fruits and its original host plant was the native North American hawthorn. However, with the introduction of domesticated apple in North America, some populations of *R. pomonella* adapted to this tree as their new host. Both hawthorn and apple *R. pomonella* races can geographically co-exist, but disparate mating behaviour between races establishes a barrier that significantly reduces gene exchange between them. Specific regions of the genome are highly divergent between these races, and some of these are thought to include multiple loci associated with adaptation to the new apple host.

Although the heterogeneity of genomic patterns of divergence is well reported in several cases of recent divergence with past or present gene flow, the significance of genomic islands in the speciation process and the mechanisms by which they are formed is debatable (White *et al.* 2010; Turner & Hahn 2010; Nosil *et al.* 2012; Cruickshank & Hahn 2014). In fact, regions of divergence might form by chance alone: if two populations become partially or totally isolated, such that gene flow is low between them, different haplotypes might become fixed by drift. This results in elevated divergence because the randomly fixed haplotypes are not shared between populations; however, the mean divergence across the genome is low because the populations still share ancestral polymorphisms in most other parts of the genome. For example, in the cited case of *Heliconius* butterflies, there is a correlation between mean divergence across the genome and the geographic distance at which two populations are from each other (Nadeau *et al.* 2013; Martin *et al.* 2013). This suggests that some divergence islands between populations that are far away from each other (i.e. there is little gene flow between them) might not harbour loci under divergent selection, but rather represent differences that were established by drift in each population. Thus, the knowledge of the genetic loci that are under divergent selection is important for our interpretation of the divergence patterns across the genome. Indeed, the *Heliconius* case is an example where some of the key loci and cis-regulatory elements controlling wing patterning are known and shown to co-localize regions of high divergence in the genome (Reed et al. 2011; Pardo-Diaz and Jiggins 2014).

Even when genomic islands are formed due to selection, different models can explain their formation, which might involve little or no role for gene flow (Noor & Bennett 2009; Cruickshank & Hahn 2014). For example, if two populations have recently become isolated from each other, they are expected to share many polymorphisms and thus divergence across the genome is low. However, different mutations might be selected in the different populations (e.g. if they are locally adaptive or if there is purifying selection against deleterious mutations), leading to a reduction in variability around the selected locus (there is a “selective sweep”). This selective sweep increases the relative divergence between the two populations around the selected locus, forming an island of divergence without any role for gene-flow (Cruickshank & Hahn 2014). In fact, genomic islands might also be formed by selection of a mutation that is advantageous in all populations, but that did not spread across the entire species range (e.g. insecticide resistance in the Malaria-transmitting mosquitoes *Anopheles gambiae*; Clarkson et al. 2014).

When it is shown that divergence islands harbour loci that have undergone divergent selection, their frequency and size will depend on how the mutations became established, how much gene flow there is between populations, and also how certain features of genomes, such as variable rates of recombination, affect the size of these islands (Feder & Nosil 2010; Feder *et al.* 2012; Nosil & Feder 2012; Flaxman *et al.* 2013). Selection and recombination have opposite effects in shaping islands of divergence. Selection on a particular locus increases the divergence levels in the locus itself, but also of the surrounding regions, due to physical linkage. In other words, neutral physically linked polymorphisms can “hitchhike” along with a selected polymorphism, becoming part of the divergence island (see “red” and “green” loci in Figure 1.1). Recombination, on the other hand, reduces the size of these divergence islands by uncoupling the association between the selected polymorphism and neutral polymorphisms linked to it. However, if a locus under selection occurs within a region of low recombination (e.g. near the centromere or within an inversion), then the region of divergence might extend for several cM around that locus, even though no other loci are under selection. Indeed, it has been often observed that regions of low recombination harbour larger islands of divergence (Feder *et al.* 2003; White *et al.* 2010; Renaut *et al.* 2013). Often, it is not clear if these low recombination regions contain loci under divergent selection or if they were simply established by chance, with no role in establishing a reproductive barrier between populations ("incidental islands"; Turner and Hahn 2010).

In summary, islands of divergence might be formed both with and without a role for selection and gene flow. Distinguishing between these can be helped by studying allele frequency changes and divergence patterns across hybrid zones, because gene flow is expected to differ between alleles in neutral loci and loci under divergent selection.

1.4 Hybrid zones

Hybrid zones provide a unique opportunity to study the genetic basis of reproductive barriers due to the contrasting behaviour of neutral and selected loci (Barton & Hewitt 1985; Barton & Gale 1993). Alleles in a neutral locus are expected to flow freely between two hybridizing populations, as long as the hybrids are fertile and produce some progeny. As a result, any differences in allele frequency that might initially exist as two populations come into contact will quickly disappear and the neutral alleles will become shared between populations (black line in Figure 1.2A). Conversely, alleles in divergently selected loci are impeded from introgressing from one population to the other, due to a reduced fitness of hybrid individuals (red line in Figure 1.2A). This creates a barrier to the exchange of such alleles, resulting in an allele frequency gradient across a geographic region that separates the two populations: such a gradient is called a "cline". A cline might also be detected as a phenotypic (rather than allelic) transition from one population to the other. In this case, the phenotypic cline is assumed to coincide with allelic clines in the loci that produce the phenotype.

Conceptually, the contrasting patterns of introgression between neutral and selected loci across a hybrid zone might produce the aforementioned heterogeneous pattern of divergence across the genome (Figure 1.2B-C). In other words, loci under selection (showing sharp clines) would fall in regions of higher divergence between the parental populations (genomic islands), whereas neutral alleles that are freely exchanged between the two populations would form the lower divergence regions of the genome (the flat "sea" of divergence).

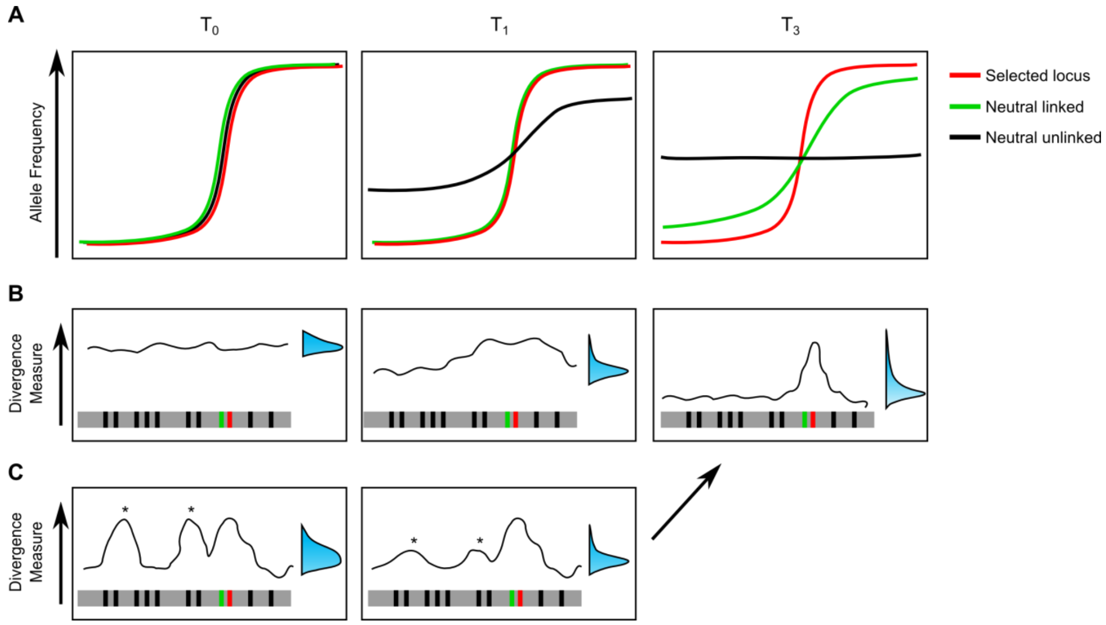


Figure 1.2 – Variation of clines and population divergence in a hybrid zone over time. The schemes conceptually illustrate a situation where two divergent populations come into contact and hybridize. The initial contact time point (T_0) and two other time points after that are represented (each step represents several generations).

A) Schematic allelic clines for three loci: a locus under selection, a neutral locus that is unlinked to it (or loosely linked), and a neutral locus that is tightly linked to it. Over time, a sharp cline is maintained for a selected locus (red line), but disappears for an unlinked neutral locus (black line). Due to physical linkage, clines in neutral loci linked to a selected locus (green line) will take longer to disappear, because it takes longer for recombination to break any associations between selected and neutral alleles.

B) Schematic of the level of nucleotide divergence across the genome, between two populations flanking a hypothesised hybrid zone. The scheme is similar to Figure 1.1. The divergence between the populations starts high across the entire genome (T_0), but progressively becomes lower around neutral loci (T_1 and T_2) due to mixing of alleles as explained in A).

C) Similar to B), but in this case the populations start off as having a highly heterogeneous pattern of divergence, where some islands of divergence are due to the fixation of alternative haplotypes that do not affect reproductive isolation ("incidental islands", indicated with an *; Turner and Hahn 2010). Over time these "incidental islands" might disappear due to gene flow, if they do not harbour loci that contribute to the reproductive isolation between the populations; the island containing a locus under divergent selection, however, remains.

The sharpness of a cline reflects the outcome of two opposing forces: dispersal and selection. Stronger selection leads to sharper clines, whereas higher dispersal widens the clines (Barton & Gale 1993). Clines might be observed across many loci, particularly if they are involved in controlling the same trait that is under selection in the hybrid zone. Although selection promotes the maintenance of such coincident clines, gene flow and recombination in the hybrid zone counteract this, by breaking any fit allelic combinations that might occur in the parental populations (Barton & Gale 1993). Therefore, the coupling of co-adapted variation depends on a balance between selection (which strengthens it) and gene flow and recombination (which act against it). In some cases, recombination might be intrinsically reduced, for example, if the co-adapted loci are physically linked or lie in a chromosomal rearrangement that impairs meiotic recombination. However, this is not always required, and fit allelic combinations between unlinked loci can be kept at high frequencies (i.e. occur in linkage disequilibrium) across the hybrid zone if selection is strong enough (Barton & Hewitt 1985).

Hybrid zones can also be used as a natural resource to map loci responsible for the traits that differentiate the parental populations that hybridize (Buerkle & Lexer 2008; Crawford & Nielsen 2013). Assuming that the hybrid zone is old enough, hybrid individuals will carry a “mosaic” genome, with interspersed portions from either parental population. This allows the detection of molecular markers that are significantly associated with certain trait states. For example, this method was used to map QTL associated with leaf morphology in hybridizing *Populus* species (Lindtke *et al.* 2013) and successfully pinpointed previously known loci that control wing colour and morphology in *Heliconius* butterflies (Nadeau *et al.* 2014).

In conclusion, the study of reproductive barriers that reduce gene flow between hybridizing populations gains from several approaches. First, it is necessary to identify the genetic basis of reproductive barriers; how many loci and how do they interact with each other? Second, we need to characterize the patterns of divergence across the genome; is divergence heterogeneous, forming discrete divergence "islands"? Finally, by studying genetic clines in hybrid zones we may answer: does selection maintain coincident clines across multiple loci that establish a reproductive barrier, or does recombination break fit allelic associations? Rarely, though, are all these aspects looked at in parallel to provide a unified picture of the nature of reproductive isolation in natural populations. *Antirrhinum majus* may prove to be an ideal model to bridge this gap, since hybrid zones are formed between polymorphic subspecies and ample genetic and emerging genomic tools are available in this system.

1.5 Study system: an *Antirrhinum* hybrid zone

The genus *Antirrhinum* includes 20 to 27 European species and/or subspecies, depending on the taxonomic convention considered (Vargas *et al.* 2009; Wilson & Hudson 2011). These species are found mostly in the Mediterranean region, occupy diverse habitats and have an extensive variability in traits such as the shape and size of leaves and flowers as well as flower colour (Schwinn *et al.* 2006; Feng *et al.* 2009; Wilson & Hudson 2011).

Despite the phenotypic diversity in this genus, *Antirrhinum* species are inter-fertile, suggesting they have recently diverged from a common ancestor and so no postpollination reproductive barriers have evolved in this genus. In fact, extensive hybridization is thought to be a part of the divergence history of *Antirrhinum*, due to non-resolvable species phylogenies using both plastid and nuclear markers (Vargas *et al.* 2009; Wilson & Hudson 2011).

Antirrhinum flowers have a characteristic morphology that is related to their insect-mediated pollination. The flowers have bilateral symmetry, with the corolla consisting of two dorsal petals, two lateral petals and one ventral petal (Figure 1.3A). The petals can anatomically be separated in two parts (Figure 1.3B): at the base of the flower, the five petals are fused to form a tube, whereas distally they form lobes (the two dorsal petals form the upper lobes, whereas the lateral and ventral petals form the lower lobes). The reproductive organs are enclosed within the corolla, making them accessible only by physically moving apart the lobes, which are shut in a spring-like manner. Pollination is carried out by large insect pollinators, in particular large bee species (e.g. *Bombus* spp. and *Xylocopa* spp.), which enter the flower to access the pollen and the nectar that accumulates in the lower part of the corolla tube (Whitney & Glover 2007; Vargas *et al.* 2010).

Although *Antirrhinum* flowers are hermaphroditic (having both male and female reproductive organs), wild *Antirrhinum* species have a physiological self-incompatibility system that impedes self-fertilization (Xue *et al.* 1996). Therefore, they are outcrossing species that depend on their pollinators to move pollen between individuals (although self-fertilization can occasionally be achieved in the glasshouse; Lucy Copsey, pers. comm.). In this context, floral features such as flower colour patterning are thought to be an important trait for pollinator attraction in this genus (Glover & Martin 1998; Shang *et al.* 2011; Whitney *et al.* 2013).

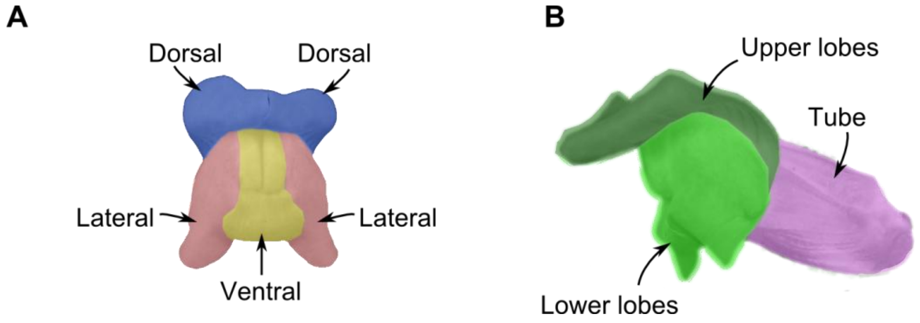


Figure 1.3 - Anatomical features of the Antirrhinum corolla.
A) A front view of the flower indicating its five petals.
B) A side view of the flower distinguishing between the tube and lobe regions of the petals. Colours were added in.

This work focuses on the study of two subspecies of *Antirrhinum majus* that inhabit the Pyrenees region: *A. m. pseudomajus* and *A. m. striatum* (Figure 1.4A). The two subspecies are very similar to each other in most visible traits, with the only conspicuous difference being their flower colour (Figure 1.4B-C). However, other unaccounted cryptic trait differences might exist between the two (e.g. floral scent; Suchet et al. 2010).

The genetics of flower colour have been extensively studied since the earliest days of genetics (e.g. Mendel's studies on peas). Partially, this is due to the genetic tractability of this trait: the loci involved are largely Mendelian and the phenotype easy to visually characterize. Flower colour in *Antirrhinum* is determined by flavonoid pigments: the yellow colour is due to aurones and the magenta-purple colour is due to anthocyanins (Geissman *et al.* 1954). The biosynthetic pathway of these pigments has been well characterized, and many of the genes involved have been cloned (Winkel-Shirley 2001). This includes genes encoding enzymes in the anthocyanin pathway as well as regulatory genes controlling them (Quattrocchio *et al.* 1999; Schwinn 1999; Schwinn *et al.* 2006; Shang *et al.* 2011; Yuan *et al.* 2013).

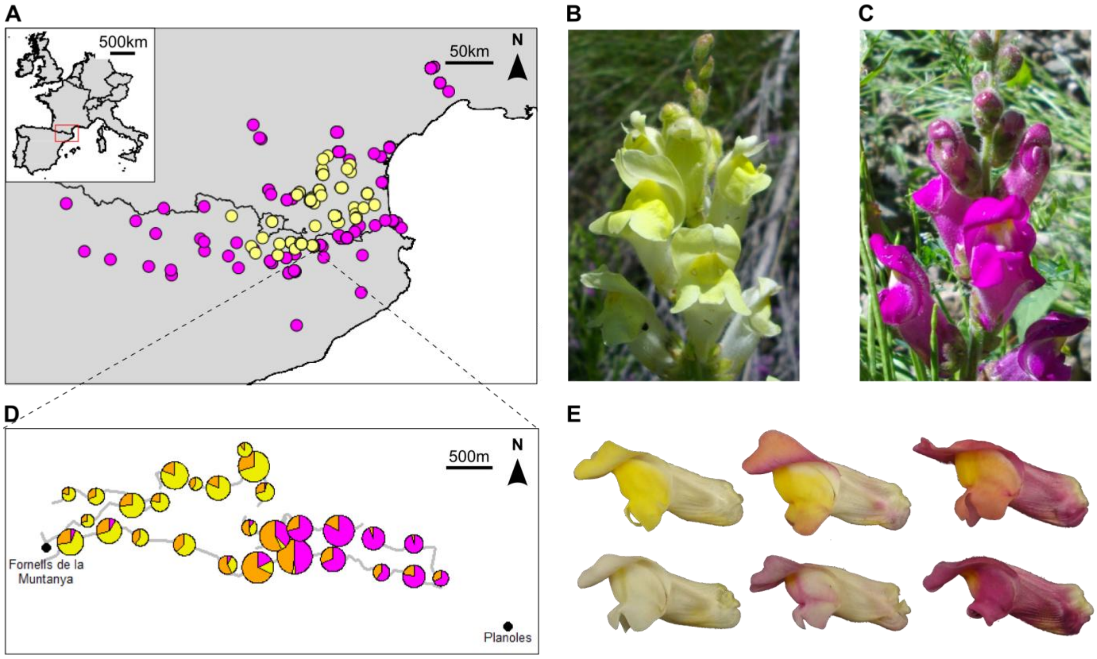


Figure 1.4 – Populations of *A. majus* subspecies and their flower colour phenotypes.

A) Distribution of *A. m. pseudomajus* (magenta circles) and *A. m. striatum* (yellow circles) populations in the Pyrenees region.

B) An *A. m. striatum* individual growing in its natural habitat.

C) An *A. m. pseudomajus* individual growing in its natural habitat.

D) Pie-charts showing the frequency of *A. m. pseudomajus* (magenta), *A. m. striatum* (yellow) and hybrid (orange) flower colour phenotypes across a hybrid zone between the two subspecies. The size of the pie-charts is proportional to the number of samples in that location. Grey lines mark the main roads along which *A. majus* grows. Two villages neighbouring the hybrid zone are also indicated.

E) Photographs of flowers from individuals from the hybrid zone shown in panel D. Examples include a *striatum*-like phenotype (top left), a *pseudomajus*-like phenotype (bottom right) as well as several hybrid phenotypes (all others). The magenta pigment can be fully spread (right), restricted within the lobes (middle) or mostly absent (left). Combined with this, the yellow pigment can be spread (top) or restricted (bottom) in the flower lobes.

In the *A. m. pseudomajus* and *A. m. striatum* subspecies, a single locus, named *SULFUREA* (*SULF*), is thought to be largely responsible for the yellow difference between the two (Whibley *et al.* 2006). *A. m. pseudomajus* carries a dominant *SULF* allele that restricts yellow pigmentation to a region where the lateral and ventral petals meet, whereas *A. m. striatum* carries a recessive *sulf* allele that allows the pigment to spread across the entire petal lobes. The difference in magenta anthocyanin pigmentation is mostly determined by two tightly-linked loci, named *ROSEA (ROS*) and *ELUTA* (*EL*) (hereon referred to as *ROS-EL*). The two subspecies carry different alleles in each of these loci, which results in spread magenta pigmentation in *A. m. pseudomajus* and hardly visible pigmentation in *A. m. striatum* (the full genetic details of *ROS* and *EL* will be considered in chapter 4).

From the three loci that establish the major difference in colour between *A. m. pseudomajus* and *A. m. striatum*, only *ROS* has been cloned. It includes a tandem duplication of two myeloblastosis (*MYB*)-like transcription factors named *ROS1* and *ROS2* (Schwinn *et al.* 2006). These genes have two repeats of the conserved *MYB* DNA-binding domain, thus belonging to the *R2R3-MYB* family of transcription factors, one of the largest in plants (Stracke *et al.* 2001).

For the most part, *A. m. pseudomajus* and *A. m. striatum* occur in isolated populations, but at the edges of each subspecies' distribution range there are zones of contact where they form hybrid zones (Whibley *et al.* 2006). One of these hybrid zones, on which this work focuses on, has a sharp cline for flower colour across two road transects approximately 2km long (Figure 1.4D). On the Eastern side of the hybrid zone there are mainly magenta-flowered individuals (*A. m. pseudomajus*) and on the Western side, yellow-flowered individuals (*A. m. striatum*) predominate. In the hybrid zone centre, a diverse array of flower colour phenotypes is found (Figure 1.4E), resulting from the segregation of alleles from the two subspecies in the major loci responsible for the subspecies' phenotypic difference (*SULF* and *ROS-EL*). The sharp phenotypic transition across this zone suggests that flower colour is under selection, since hybrid phenotypes have not significantly introgressed to either parental population. Supporting this view is the fact that, along with the phenotypic cline, there is a correlated allelic cline for the *ROS1* gene, while presumably neutral loci have no significant clines across the hybrid zone (Whibley *et al.* 2006).

Because flower colour is a genetically tractable trait and several mutant lines are available in *A. majus*, this is an ideal system to genetically dissect a trait that reduces gene flow between hybridizing populations. This present work focuses on the linked *ROS-EL* loci that control the magenta pigment of *Antirrhinum* flowers. Although *ROS* has already been characterized at a molecular level (and the *ROS1* gene studied in the hybrid zone), the linked *EL* locus remains to be identified. These loci are interesting because they genetically interact to produce the magenta phenotype of the flowers, providing an opportunity to investigate how the *A. m. pseudomajus* and *A. m. striatum* allelic combinations are maintained despite gene flow in the hybrid zone. I will consider the impact of selection and gene flow on the observed genomic pattern of divergence between *A. m. pseudomajus* and *A. m.*

striatum in the *ROS-EL* region (chapter 3). This approach is combined with genetic and molecular experiments that allowed the fine-mapping of individual loci controlling different aspects of the flower colour phenotype (chapters 4 and 5). Finally, I will consider the phenotypic consequences of recombination between the mapped loci in the *pseudomajus* x *striatum* hybrid zone (chapter 6). I will discuss the main findings by considering how the effects of selection, gene flow and recombination shape the genomic divergence patterns between tightly-linked loci that, together, contribute to a reproductive barrier between hybridizing populations (chapter 7).

2 Materials and Methods

2.1 Plant Material from wild populations

In the field, the following data were obtained from wild *A. majus* plants: a global positioning system (GPS) coordinate; a flower colour score; a photograph of a representative open flower; and leaf material (for DNA extraction). Each individual was tagged with a unique code that consists of a letter specific for each year of sampling (in this work J-M for years 2010-2013), followed by a four-digit number (unique to each individual). The collection occurred between the months of May-July (occasionally extending to August), for individuals with open flowers only.

Each sampled individual's coordinates were collected using a GPS device (Trimble) with a mean accuracy of ~1m.

The magenta and yellow colour of flowers was scored according to Whibley (2004) (magenta score detailed in chapter 4).

One flower from each individual was photographed against a black background, using a digital camera (Nikon Coolpix 995 or Olympus XZ1). These photographs were used for later confirmation of flower scores. Light conditions involved a mix of indoor natural light and the use of incandescent/halogen light bulbs illuminating the flowers. These conditions were not standardized across the different years of sampling and therefore characteristics of the colour such as hue, brightness and saturation were variable between photographs. For this reason, these photographs were used only to qualitatively confirm certain phenotypic scores with regards to the presence/absence of pigment in certain regions of the flowers.

Leaf material was collected from each individual, usually consisting of 4 young leaves (about 1cm long) and 2 older leaves (about 2cm long). In some cases, less than this amount of material was collected, due to a less vigorous vegetative condition of some individuals. Fresh leaves were stored in individual glassine envelope bags, which were placed within a plastic bag containing silica gel (Fisher Scientific) for drying the leaf tissue. This allowed long-term storage of dry leaf material for later DNA extraction in the lab.

The sampling transect consisted of two main roads that extend for 5-6km. This included the central region of the hybrid zone as well as its flanks (Figure 1.4D). Because the sampling season extended for several weeks, this transect was surveyed several times per season, ensuring that the majority of flowering individuals was sampled: this amounts to ~2000-3000 samples per year.

2.2 Plant material from glasshouse experiments

2.2.1 Nomenclature for individuals grown in the glasshouse

Plants grown in the glasshouse are named with a letter that sequentially changes between sowing seasons (there are generally two sowing seasons per year, approximately between OCT-APR and APR-SEP), followed by a number unique to each family. A family is considered to be the progeny of a self pollination or a particular cross and is usually derived from a single capsule (an *Antirrhinum* flower gives a single capsule with ~50-200 seeds). Each individual within a family has a unique number, shown after a hyphen. For example Y207-09 refers to individual number 9 from family number 207 sown in the autumn of 2010 (letter Y).

Some family numbers are reserved to particular stock lines, which have been kept at the John Innes Centre for several years. In this case, the stock name or number is used. For example, in this work I will often refer to stock line *JI7*, and the mutant line *rosea^{dorsea}* (detailed in chapter 4). These lines are highly introgressed and will therefore be assumed to be homozygous for all loci in the genome.

2.2.2 Growth conditions and crosses

Plants were grown in the John Innes Centre glasshouses under conditions that varied slightly depending on the sowing season. During the autumn/winter months (Oct-Mar) plants were kept under artificial light conditions of daily cycles of light and dark of 8h and 16h, respectively. During the spring/summer months (Apr-Sep) plants were kept at ambient temperature and light, either within the glasshouse (for smaller numbers of plants or for plants to be crossed) or outside on benches open to natural weather conditions (for sowing large numbers of plants, e.g. the recombinant screens detailed in chapter 4).

Crosses were always performed indoors, by emasculating flower buds at an early stage of development, before the anthers were mature. Two to three days after emasculation, when the flower had opened and the pistil matured, pollen from a donor parent was then deposited on the stigma of the emasculated flower. A cross was considered successful if a swollen capsule started to form, which would then be left to fully mature on the plant, until it was about to open. Collected seeds were stored in glassine envelope bags until needed.

2.3 DNA extraction methods

Several methods were used for DNA extraction, depending on how many samples had to be processed and the quality of extraction required. For long term storage, higher quality DNA was required (using a slower and/or costly method of extraction). On the other hand, DNA samples only needed for a few genotyping reactions (e.g. for genotyping segregating families or to aid in setting up crosses) were extracted with faster and/or cheaper extraction methods that yielded poorer quality DNA.

2.3.1 Large scale, high quality DNA extraction

For extraction of large numbers of samples for which high-quality DNA was required (e.g. for hybrid zone samples), the DNeasy 96 Plant Kit (QIAGEN) was used. Extractions were done by Richard Goram who provides a DNA extraction service at the John Innes Centre. Usually, 2 young leaves (frozen or silica-dried; ~100 mg of fresh weight) were used per sample and the DNA extracted following the manufacturer's protocol. The DNA was resuspended in 200µl of AE buffer (QIAGEN).

All of the samples from the hybrid zone were extracted using this method. The amount and state of this leaf material were highly variable (due to the variable condition of the individuals in the field) and so the final DNA concentration varied between ~5 - 40 ng/µl.

2.3.2 Small scale, high quality DNA extraction

For extraction of lower numbers of samples with high quality DNA, the DNeasy Plant Mini Kit (QIAGEN) was used, following the manufacturer's protocol. Depending on the amount of starting material (between 1-2 young leaves), the final DNA was resuspended in 50 or 100µl of AE buffer (QIAGEN).

2.3.3 Large scale, low quality DNA extraction

For extraction of lower quality DNA from a large number (hundreds) of samples, an in-house method adapted from Green (2007) was used. This method allows processing two batches of 96 samples in each extraction, by using racks of 96 × 1.2ml Collection Microtubes (QIAGEN) and standard 96-well PCR plates (200µl volume).

Each of the 96 collection microtubes (QIAGEN) in a rack was loaded with one young leaf (~1cm long) per sample. A 3mm tungsten carbide bead (QIAGEN) was added to each tube, lids were added to seal the tubes and the tissue disrupted in a TissueLyser (QIAGEN) machine by using a 30 second shaking step at 20Hz. After disruption, beads were removed by removing the lids and inverting the tubes, taking care that most of the plant tissue remained attached to the tube walls. 200µl of extraction buffer [100mM Tris (pH 8.0); 1.4M NaCl; 20mM EDTA (pH 8.0); 2%

(w/v) CTAB] was added to each tube in the rack, sealed and then shaken vigorously by hand. The tube rack was incubated for 20-30 minutes at 55^{o}C in a laboratory oven, and shaken 2-3 times during this period. After incubation, samples were centrifuged for 1 minute at 4000rpm in a Sigma 4K15 centrifuge (with a 4 plate rotor) to remove samples from the walls of the tube. 100µl of chlorophorm was added to each tube, and samples mixed thoroughly. Samples were centrifuged for 5 minutes at 4000rpm in a Sigma 4K15 centrifuge. 120µl of the supernatant was transferred to a 96-well PCR plate, to which 80µl of isopropanol had been previously added. Samples were mixed by pipetting. The PCR plate was sealed and centrifuged for 10-15 minutes at 4000rpm in a Sigma 4K15 centrifuge (after this step a white DNA pellet was visible). The supernatant was discarded by carefully inverting the PCR plate. The DNA pellets were washed by adding 180µl of 70% (v/v) ethanol to each well, which was then discarded by carefully inverting the plate. Samples were left to air dry for several minutes, until the DNA pellet became transparent. DNA was ressuspended in 50µl TE buffer [10mM Tris-HCl (pH8.0); 1mM EDTA(pH8.0)] and stored at -20^{o}C until use.

2.3.4 Small scale, low quality DNA extraction

For extraction of lower quality DNA from a small number (dozens) of samples, an in-house method following Green (2007) was used.

One young leaf (~1cm long) per sample was placed in a 1.5ml Eppendorf tube. Tissue was disrupted manually by using a micropestle (chilled in liquid nitrogen if the leaf material was frozen). 400µl of extraction buffer [100mM Tris (pH 8.0); 1.4M

NaCl; 20mM EDTA (pH 8.0); 2% (w/v) CTAB] was added to each tube and vortexed vigorously. The tube was incubated at 65^{o}C for 25-30 minutes, and shaken 2-3 times during this period. Samples were left to cool for 2-3 minutes. 200µl of chlorophorm was added and the tube vortexed vigorously. The sample was centrifuged for 5 minutes at 12000rpm in a microcentrifuge. 300µl of supernatant was transferred to a new tube, 200µl of isopropanol added, and mixed by inverting the tube several times. The sample was centrifuged for 10 minutes at 12000rpm in a microcentrifuge (after this step a white DNA pellet was visible), and the supernatant discarded. The DNA pellet was washed by adding 500µl of 70% (v/v) ethanol, which was then discarded by carefully inverting the tube. Samples were left to air dry for several minutes, until the DNA pellet became transparent. DNA was ressuspended in 50µl TE buffer [10mM Tris-HCl (pH8.0); 1mM EDTA(pH8.0)] and stored at -20^{o}C until use.

2.3.5 CTAB DNA extraction from pooled leaf tissue from the hybrid zone

A CTAB-based method of DNA extraction was used for medium-sized extractions of leaf tissue from pools of plants sampled from the hybrid zone (chapter 3). These extractions were made by Desmond Bradley (Coen Lab, JIC). The leaf material used in these extractions was collected from 50 or 52 randomly chosen plants located at different distances from the hybrid zone centre. Half of the leaves from each individual were kept separate (for individual DNA extractions) and the other half was pooled together (for a pooled DNA extraction). The leaf samples were silica-dried, both from individuals and pools.

The pooled leaf material was ground with a mortar and pestle (because the material was dried, grinding was done at room temperature). 5-7mg of tissue powder was added to a 15ml corning tube containing 5ml of DNA extraction buffer [100mM Sodium diethyldithiocarbamate; 10mM EDTA (pH 8.0); 3x SSC (450mM sodium chloride; 45mM trisodium citrate)] and 1.25ml 10% (w/v) SDS. After thoroughly mixing, 4ml chloroform was added to the tube and mixed. The tube was left for 10min, with occasional mixing. The sample was centrifuged for 10min at 3000rpm in a Sorvall RC3C centrifuge. The aqueous phase was transferred to a new 15ml corning tube and 3.2ml of phenol added. The sample was left for 10min (with occasional mixing), and then 3.2ml of chloroform was added and left again for 5min (with occasional mixing). The sample was centrifuged for 10min at 3000rpm in a Sorvall RC3C centrifuge and the aqueous phase (~6ml) transferred to a clean 15ml corning tube. The tube was filled to the top mark with 100% ethanol (~9ml) and mixed until a DNA precipitate was formed. The DNA was pelleted by centrifugation for 5min at 3000rpm in a Sorvall RC3C centrifuge. The supernatant was discarded and the DNA pellet mixed with 1ml TE buffer [10mM Tris-HCl (pH8.0); 1mM EDTA(pH8.0)] and 5µl RNase A (1mg/ml). The tube was left overnight at 4^{o}C to dissolve the DNA and then re-precipitated by adding 120µl 5M NaCl and 1ml of CTAB buffer [0.5M Tris-HCl (pH 7.5); 10mM EDTA (pH 8.0); 20mg/ml Cetyltrimethylammonium bromide (CTAB)]. After this step a precipitate formed, which was washed with 1ml of 70% (v/v) ethanol and 30% (v/v) 0.5M NaCl. The tube was left in the ethanol/NaCl solution for 1h, with occasional mixing. The DNA precipitate was transferred to an Eppendorf tube and left to air dry before ressuspension in 100-150µl of TE buffer by leaving overnight at room temperature. The DNA was stored at 4^{o}C until use.

2.4 RNA material

The plant tissue used for RNA extraction in this work, consisted of flower buds varying between 5-10mm long (Figure 2.1). During collection, a flower bud was cut from the plant, the sepals, stamens and pistil were removed with tweezers, and the corolla placed in a 2ml Eppendorf tube, which was immediately placed on dry ice. These manipulations were as quick as possible, to avoid RNA degradation. After collection, the material was stored at -80^{o}C until extraction.

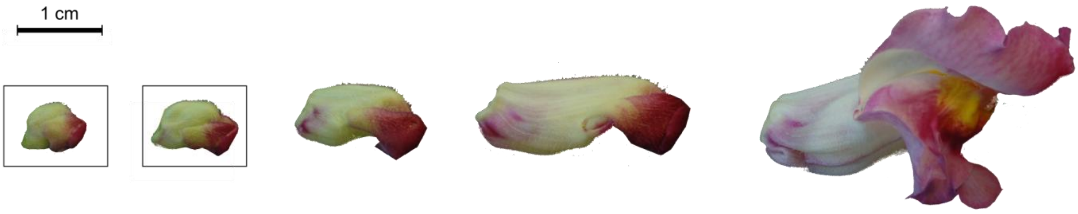


Figure 2.1 - Size series of corollas from an *Antirrhinum* flower.
For RNA extraction, corolla tissue from flower buds 5-10mm long (boxed) was used.

For RNA extraction, the corolla tissue from a single bud was placed in a 2ml Eppendorf tube pre-chilled in dry ice, which contained a 3mm tungsten carbide bead (QIAGEN). The tube was placed in a TissueLyser Eppendorf tube adapter (QIAGEN), which had been pre-chilled in dry ice to ensure that the samples remained frozen until the first step of RNA extraction. The tissue was disrupted in a TissueLyser (QIAGEN) machine by using two 30 second shaking steps at 25Hz.

The pulverized corolla tissue was used for RNA extraction using the RNeasy Plant Mini Kit (QIAGEN), following the manufacturer's protocol (the optional steps of on-column DNase digestion were performed as instructed). The RNA was resuspended in 50µl of RNase-free water and 1µl was run on a 1% (w/v) agarose gel to confirm RNA integrity. Samples were stored at -80^{o}C until use.

2.5 Genotyping

Several methods for genotyping individuals were used in this work. In some cases, individuals were genotyped by using single nucleotide polymorphisms (SNPs) from whole-transcriptome (RNAseq) data (section 2.7.2.2). All other methods used are based on the *polymerase chain reaction* (PCR) method. The primers used in these genotyping assays were all linked to the *ROS-EL* genomic region (Figure 2.2). A list of primers for each marker is given in Table 2.1, where the column named "Genotyping Method" corresponds to each of the protocols detailed in this section.

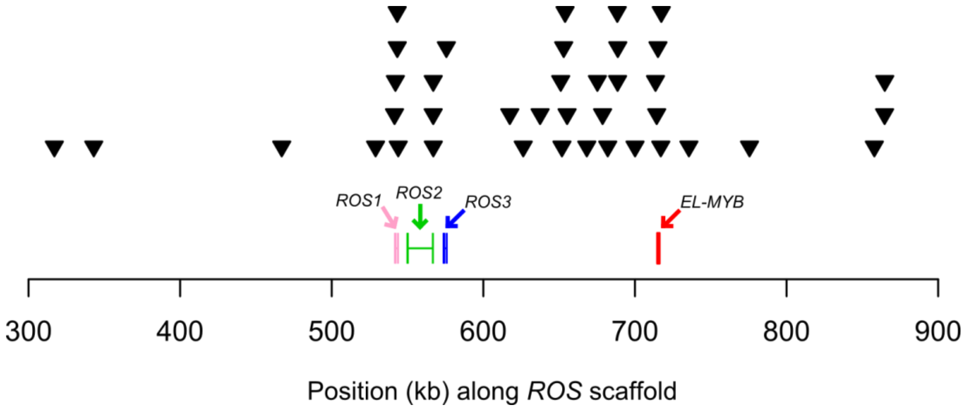


Figure 2.2 – Location of markers used in the *ROS-EL* region.

The triangles correspond to the markers in Table 2.1 and are plotted in relation to their position within the *ROS* scaffold. The coding sequence of key genes is indicated (vertical lines are exons and horizontal lines are introns).

Table 2.1 – Primers for polymorphic markers used for genotyping mapping populations and natural populations.

The primers are ordered by their location in the *ROS* scaffold. Primer numbers refer to the Coen Lab oligo database. Primer orientation is given in relation to the *ROS* scaffold. In cases where the marker targets a particular SNP, its position is indicated. Markers in the genes *ROS1, ROS2*, *ROS3* and *EL-MYB* are indicated. The marker numbers from this table are used throughout the thesis.

Marker number	Genotyping Method	Primer orientation	5' primer position	focal SNP	Gene	Primer number	Sequence (5'-3')
1	MULTIPLEX	F	316853			#1635	TTGGCCCAACTAAGATGATAAG
		R	317232			#1552	CTTACGAAACAAATCGGCTCAT
2	MULTIPLEX	F	342846			#1547	TTGGTGGGCCTAACTTTTCTTA
		R	343237			#1636	TCAACAATTCTCACCCCCTGTT

(Table 2.1 continued)

3	MULTIPLEX	F	466813			#1634	TTCTCGTCACTTTACAACACTGAAC
		R	467222			#1569	GAAACATGGGGACTTCAACAAT
4	KASP	F	528885	528910		#1480	AGGTTTCTGAAGCGCCAGGTTC
		R	528931			#1481	GAAGGTGACCAAGTTCATGCTAATGCGACAACAACGTCTAACG
		R	528931			#1482	GAAGGTCGGAGTCAACGGATTAATGCGACAACAACGTCTAACA
5	PCR	F	541186		ROS1	#1257	GGCTCCACCCTATGATGTATGT
		R	541644		promoter	#1258	GAGTACCCCTTGAGCGAAACTT
6	KASP	F	541834	542000	ROS1	#1483	TGGCATCAAGTTCCACACAGAGCAG
		R	542020		intron1	#1911	GAAGGTGACCAAGTTCATGCTCAACATTGACGTACGGTATTC
		R	542020			#1912	GAAGGTCGGAGTCAACGGATTCAACATTGACGTACGGTATTT
7	MULTIPLEX	F	543023		ROS1	#2525	ACTATCCGAGTTGAACAATCTGGCCA
		R	543395		intron2	#1181	AGTTTCAACAAGACGGGAGCTA
8	CAPS	F	542992	543323	ROS1	#1182	CAATGTGCATGTCCTTCCTAAA
		R	543503		intron2	#1247	ATGGACCCCGCTAAACACTTA
9	SANGER	F	543581		ROS1	#1754	TGTCCGGTAAGAAAGAAAAGGA
		R	544170		exon3	#1755	TCTCATTGTCTAACGGTTGCA
10	SANGER	F	566650		ROS2	#1750	GCCTAAATCCTTAGGAAATTGC
		R	567213		exon3	#1751	GGCTTAAACAATCCGTTGTGA
11	CAPS	F	566775	566852	ROS2	#1259	TTGGAATACTCATGTGGGGAAG
		R	567136		exon3	#1260	ATTCAGACATTTTTCCGGTTTG
12	KASP	F	566979	567004	ROS2	#2298	AGATTATGAGAAGCAAAAG
		R	567023		exon3	#2296	GAAGGTGACCAAGTTCATGCTGTTGAGGCCACATTATTGTG
		R	567023			#2297	GAAGGTCGGAGTCAACGGATTGTTGAGGCCACATTATTGTA
13	KASP	F	575590	575623	ROS3	#1557	GGATGGATTATCAAAATTCTAC
		R	575644		intron2	#1555	GAAGGTGACCAAGTTCATGCTCTACAAAAGATTATGTCCTACT
		R	575644			#1556	GAAGGTCGGAGTCAACGGATTCTACAAAAGATTATGTCCTACA

(Table 2.1 continued)

14	PCR	F	616606		#194	GGAGAGGAAGGGGTTGTTGG
		R	618300		#246	AGAGTTGTGGGATTGGAGTAA
15	SANGER	F	626228		#1440	AAATTAAACTAAAAACGCGAGGAT
		R	626401		#1439	TCAATATCTTTCCTACTCACGTCCT
16	MULTIPLEX	F	637333		#1637	ACGTCGAATTTGTTGAAGACCT
		R	637553		#1605	TGCAACATAACTAAATTCCCACTC
17	SANGER	F	650759		#1594	AGAAGTTTGTACCCGGAAATGA
		R	651324		#1595	GTTTTGGCTTTCTTTGAAGCAC
18	SANGER	F	651551		#1596	AGGATCTTGTCCCGAATGGT
		R	652136		#1597	AGTAGCCAAAACCTGCACAAAT
19	KASP	F	652993	653015	#2302	GAAGGTGACCAAGTTCATGCTAAATTAAGCTGTACATTAATTAC
		F	652993		#2303	GAAGGTCGGAGTCAACGGATTAAATTAAGCTGTACATTAATTAT
		R	653040		#2304	TTCAGCAGTTTAAGGGAG
20	PCR	F	653629		#1598	CCCTGTGACCTTGTCTTCTTTT
		R	654198		#1599	GAAGTCCTTTGTTTTGCTGAGA
21	SANGER	F	654805	655252	#2192	CTGGTGTTCAAGGAGTTGGTT
		R	655699		#2193	AGCAAGCAGTATCGCATCATT
22	SANGER	F	667783	668233	#2178	CATCAAAGTGGGGAAGAAGGT
		R	668683		#2179	TAAGAAAAATGGGGCAAACAG
23	SANGER	F	674804	675235	#2180	TCTGTGTGCAGGCAAGAAACT
		R	675666		#2181	GCAGCAGTAAGAAGGAACCAA
24	SANGER	F	678075	678655	#2148	TAACAAGGGCCAAAAAGAGGT
		R	679234		#2149	GGTGCCAACAACTTAAAACGA
25	SANGER	F	680007	680529	#2150	AATCGTATCTGGTGCTGATGG
		R	681051		#2151	CGCTGATCCAAGCTGATAAAG
26	RNAseq	-	-	688352	-	-

(Table 2.1 continued)

27	RNAseq	-	-	688550		-	-
28	RNAseq	-	-	688746		-	-
29	SANGER	F	699584			#1535	ATTCAGATCCAAGCATGAAAGC
		R	700542			#1536	GTGCATCACAACTCACAATGAA
30	PCR	F	712107		EL-MYB	#1888	AAACGTGAAGTAATTCTAGCTGCA
		R	715005		exon3	#1607	CGAATGGATGATGAAGTGAAGA
31	PCR	F	713692		EL-MYB	#1886	AAACTCGATCCACCTTGGTATT
		R	715005		exon3 to 3’ UTR	#1607	CGAATGGATGATGAAGTGAAGA
32	MULTIPLEX	F	715097		EL-MYB	#1640	ATGAAGAAAAAGCTTAGGTGAACT
		R	715266		intron2	#1646	GGGATGGTGTGCTACCTTTT
33	KASP	F	716969	717045	EL-MYB	#1615	CATTGTCATGACTCGTTCAACA
		R	717066		promoter	#1872	GAAGGTGACCAAGTTCATGCTGAATATATTTAAAGTGAGAGTA
		R	717066			#1873	GAAGGTCGGAGTCAACGGATTGAATATATTTGAAGTGAGAGTC
34	PCR	F	716969		EL-MYB	#1615	CATTGTCATGACTCGTTCAACA
		R	717779		promoter	#1616	TTAAACTGAAAGGCAGGCAATC
35	SANGER	F	735046			#2440	CAGACGTTTTAGGTTCCACCA
		R	736154			#2441	GGTGTTCATCCACATTGCTCT
36	SANGER	F	774993			#2438	ACTCGGAAGATGAGGGAAAAA
		R	776141			#2439	ATCAAGTCGTTGTCGTCGTTC
37	PCR	F	857452			#1511	AGTCCCCGAAATGTAAGTTGTG
		R	858371			#1512	CCGAGTCTTTCTAGCCACGTAT
38	MULTIPLEX	F	864557			#1639	CAGATTGGTTCTTACACCGTCA
		R	864794			#1531	GAAGTGGAATTTTGTGGAGGAG
39	SANGER	F	864282			#1513	GTTCGGTTTCTCGAATGGATAC
		R	865281			#1514	AGATGACAAAGGTGGCAAATCT

2.5.1 Genotyping method: PCR

A standard PCR reaction was used for genotyping indel polymorphisms that were distinguishable by running the PCR products in 1-2% (w/v) agarose gels.

The PCR reaction was prepared using QIAGEN's *Taq* DNA Polymerase, as follows:

Reagent	**20µl final volume**
Water	10.9µl
10x PCR Buffer	2µl
1mM dNTPs	2µl
5µM forward primer	2µl
5µM reverse primer	2µl
Taq polymerase (5 units/µl)	0.1µl
DNA*	1µl

* A constant volume of DNA was added, independently of the sample's concentration

PCR reactions were performed in a thermocycler using the following program:

Temperature	**Step duration**	**Number of cycles**
94 °C	5 min	1
94 °C	30 sec	35
55 °C *	30 sec	
72 °C	1 min / kb	
72 °C	5 min	1

* Usually an annealing temperature of 55°C was used but it could vary between 50-60°C in some cases

After the reaction, 2µl of 6X DNA Loading Dye was added to 10µl of the PCR product, which was analysed by agarose gel electrophoresis. Generally, a 1% (w/v) agarose gel was used, except for markers 5 and 34 in Table 2.1, where a higher percentage of 2% was used to resolve fragments (see Figure 6.3).

2.5.2 Genotyping method: KASP

KASP™ technology is a PCR-based genotyping method that uses fluorescent dyes to discriminate between bi-allelic SNPs (details of the method can be found at http://www.lgcgenomics.com/genotyping/kasp-genotyping-reagents/how-does-kasp-work/; accessed March 2014).

For this study, the assay requires the design of three primers: two primers specific for each SNP allele and a primer that is common for both alleles. The SNP-specific primers were designed such that the first 3' nucleotide complemented the target SNP (while the rest of the primer sequence was identical between those two primers). A tail was added to each of these primers at the 5' end, composed of a sequence complementary to the dye-containing FRET cassettes in the KASP Master Mix (LGC). The common primer was designed no further than 300bp away from the SNP-specific primers. A primer mix was made by mixing the three primers to a final concentration of 12µM for each of the SNP-specific primers and 30µM for the common primer.

The genotyping reaction was prepared according to the manufacturer's protocol. Reactions were performed either in 96-well or 384-well white-coloured PCR plates. For 96-well plates the final volume of reaction was 10µl, whereas for 384-well plates the final volume was 5µl.

The PCR assay was prepared as follows:

Reagent	**10µl final volume**	**5µl final volume**
Water	4µl	2µl
KASP Master Mix	5µl	2.5µl
Primer Mix	0.14µl	0.07µl
DNA	1µl	1µl

Because the method relies on a PCR-based reaction, it is relatively robust to variable DNA concentrations. Therefore, 1µl of DNA (extracted as in section 2.3) was always used, regardless of variable concentrations between samples.

The PCR reaction was performed in a regular thermocycler using the following program:

Temperature	**Duration**	**Number cycles**
94 ^{o}C	15 minutes	1
94 ^{o}C	20 seconds	10
Touchdown over 65-57 ^{o}C (dropping 0.8 ^{o}C per cycle)	60 seconds	
94 ^{o}C	20 seconds	30 - 40
57 ^{o}C	60 seconds	

The number of cycles in the last steps of the PCR reaction varied between 30 and 40, depending on the intensity of the fluorescence signal after the reaction (this can vary between primers). If the signal was not strong enough after 30 cycles, the program would be resumed for 10 additional cycles (40 cycles in total).

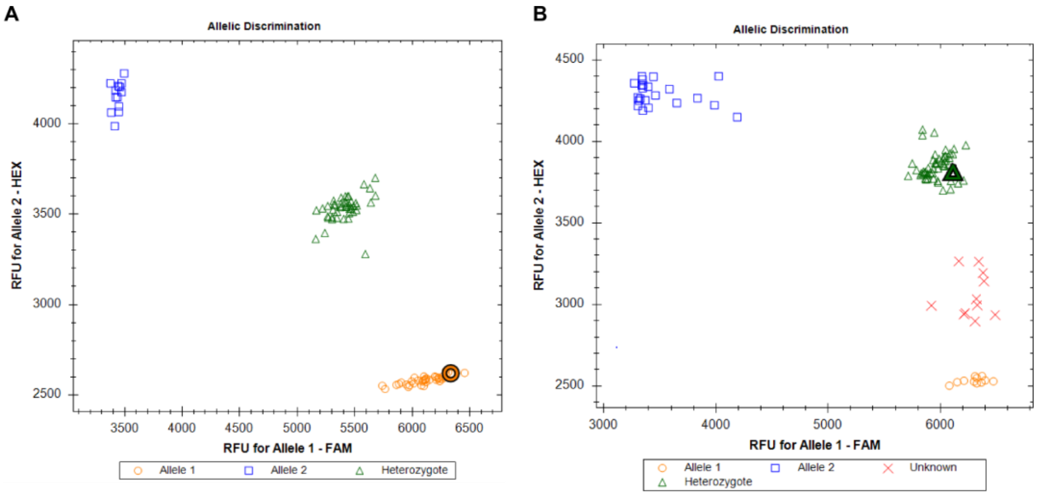


Figure 2.3 - Examples of KASP genotyping result.

The figures show a snapshot from the "Allelic Discrimination" tool in the BioRad CFX Manager 3.0 software. The software plots "Relative Fluorescence Units" for each fluorescent dye, in this case FAM (x-axis) and HEX (y-axis), which are the dyes used in the KASP assay. Samples are expected to form three clusters, corresponding to two homozygous genotypes (blue squares and orange circles) and one heterozygous genotype (green triangles). A reliable genotyping is based on the formation of tight clusters of points in the plot.

A) Example of 96 samples genotyped with a marker in the *ROS1* gene (marker no. 6 in Table 2.1). Three clear clusters are visible in this assay.

B) Example of 96 samples genotyped with a marker in the *EL-MYB* gene (marker no. 33 in Table 2.1). In this case, some samples are not clearly clustering as homozygous or heterozygous and thus are marked as "unknown" genotype (red crosses).

After the PCR reaction, the fluorescence for each well was read in a Real-Time PCR thermocycler (BioRad's *CFX96* was used for 96-well plates and Roche's *LightCycler 480* was used for 384-well plates). The fluorescence was read for FAM and HEX dyes and sample genotypes were determined using each machine's software: the "Endpoint Genotyping" tool was used on the *LightCycler 480* software (version 1.5.0); the "Allelic Discrimination" tool was used on the *Bio-Rad CFX Manager* software (version 2.1).

The fluorescent values for the FAM and HEX dyes from each sample were plotted against each other and each sample's genotype was scored according to their position on this plot (explained in Figure 2.3).

2.5.3 Genotyping method: Sanger sequencing

The Sanger DNA sequencing method was used to genotype SNPs and small indels from regions amplified by PCR (using the protocol in section 2.5.1). The amplified regions were between 500bp - 1kb.

The sequencing PCR reaction was prepared using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies), as follows:

Reagent	**10µl final volume**
Water	5µl
Ready Reaction Premix	2µl
BigDye sequencing buffer	1µl
5µM primer	1µl
PCR product*	1µl

* The product from a PCR reaction was used directly without any purification step

The sequencing PCR reaction was performed in a thermocycler using the following program:

Temperature	**Step duration**	**Number of cycles**
96 °C	30 sec	25
55 °C	30 sec	
60 °C	1 min 30 sec	
60 °C	10 min	1

The PCR product was sequenced using the “ready-reaction” sequencing service provided by TGAC (Norwich) or Eurofins MWG Operon (Germany). Identification of SNP and small indel polymorphisms was made using the *Mutation Surveyor* software (demo version 4.0).

2.5.4 Genotyping method: multiplex PCR

A multiplex PCR was used to combine seven markers in a single genotyping reaction (markers 1, 2, 3, 7, 16, 32 and 38 in Table 2.1). These markers were analysed by capillary electrophoresis, which allowed the separation of fragments differing by as little as 3bp in size (due to indels or microsatellites). The seven combined markers were distinguished from each other based on the fragment size range produced by each. Also, one of the primers from each marker was labelled with a fluorescent dye (6-FAM, VIC or NED), which further allowed the distinction between markers producing fragments in a similar size-range as they had different dye colours.

First, a primer mix was prepared containing all 14 primers (2 from each marker) at a final concentration of 2µM each.

The PCR reaction was prepared using the KAPA2G Fast Multiplex PCR Kit (KAPPA Biosystems) or the QIAGEN Multiplex PCR Kit (QIAGEN). In both cases, the reaction was prepared as follows:

Reagent	20µl final volume
Water	7µl
2x Master Mix (QIAGEN or KAPA2G)	10µl
10x Primer mix (each primer at 2µM)	2µl
DNA*	1µl

* A constant volume of DNA was added, independently of the sample's concentration

The PCR reaction was performed in a thermocycler using the following program:

Temperature	Step duration	Number of cycles
95 °C	3 min	1
95 °C	15 sec	28
60 °C	1 min	
72 °C	45 sec	
60 °C	30 min	1

After the reaction, 1µl of the PCR product was mixed with 10µl Hi-Di Formamide (Applied Biosystems) and 0.1µl GeneScan-500 ROX size standard (Applied Biosystems). This was analysed by fluorescent capillary electrophoresis by Richard Goram, who provides such a genotyping service at the John Innes Centre. Genotypes were scored using the *GeneMarker* software (demo version 2.6.0).

2.5.5 Genotyping method: CAPS

Cleaved Amplified Polymorphic Sequences (CAPS) markers allow the distinction of a SNP by differences in restriction fragment sizes from a PCR amplicon of a target region (Konieczny & Ausubel 1993). In this work, two CAPS markers were used (markers 8 and 11 in Table 2.1). In each case, a standard PCR reaction was performed as detailed in section 2.5.1.

The PCR product was then used in a restriction digestion reaction, prepared as follows:

Reagent	10µl final volume
Water	2.8µl
10x Buffer*	1µl
Restriction enzyme**	0.2µl
PCR product	6µl

* The restriction buffer used depended on the restriction enzyme.
** For marker 8 SpeI (Roche); for marker 11 MseI (Roche)

The samples were incubated in a water bath for 2h at 37^{o}C. After incubation, 2µl of 6x DNA loading dye was added to each reaction product, which was analysed by electrophoresis in a 2% (w/v) agarose gel.

2.6 *ROS1* BAC clone

A bacterial artificial chromosome (BAC) clone consisting of ~200kb of *A. majus* genomic sequence and containing the *ROS1* gene was sequenced (using *Illumina* technology) to aid in the assembly of the genomic sequence surrounding this gene (detailed in chapter 3). This BAC was part of a library provided by Zsuzsanna

Schwarz-Sommer (Causier *et al.* 2010) and screened with a probe for the *ROS1* gene by Xianzhong Feng (Coen Lab, JIC). The BAC vector was *pIndigo-BAC5*, and the frozen stock number for the *E. coli* clone is “Coen glycerol #2349”.

2.6.1 Bacteria growth conditions

A sterile toothpick was used to stab the top of the frozen glycerol stock and the bacteria spread over an LB agar plate with 25µg/ml chloramphenicol. Bacteria were left to grow overnight at 37^{o}C. A single colony was picked from the plate and inoculated in a culture of 10ml liquid LB medium with 25µg/ml chloramphenicol. This culture was grown overnight at 37^{o}C with vigorous shaking. This culture was used for DNA extraction of the BAC plasmid.

2.6.2 BAC DNA extraction

Cells from the liquid culture were harvested in a 2ml Eppendorf tube, by centrifuging 30s at 12000rpm in a microcentrifuge. This step was repeated three times, such that 6ml of the cell culture were used in total.

DNA was extracted using the QIAprep Spin Miniprep Kit (QIAGEN) with a modified protocol. The initial steps of extraction followed the kit’s protocol: the cell pellet was resuspended in 250µl buffer P1, followed by the addition of 250µl buffer P2 and mixed; 350µl buffer N3 was added and mixed immediately; the tube was centrifuged for 10 min at 12000rpm in a microcentrifuge. After this step, the protocol no longer followed the kit’s instructions. The supernatant was transferred to a new tube and 0.6 volumes isopropanol added and mixed by gently inverting the tube several times. The DNA was pelleted by centrifugation for 10 min at 12000rpm in a microcentrifuge. The supernatant was discarded and the pellet washed by adding 1.5ml 70% (v/v) ethanol and leaving to rest for 15 min. The liquid was poured off and 0.5ml 70% (v/v) ethanol added to wash briefly and then discarded. The DNA pellet was brought to the bottom of the tube by centrifugation for 1 min at 12000rpm in a microcentrifuge. Any excess liquid was removed with a pipette and the DNA pellet left to air dry until transparent. The DNA was resuspended in 30µl TE buffer [10mM Tris-HCl (pH8.0); 1mM EDTA(pH8.0)] and stored at -20^{o}C until use.

2.6.3 BAC sequencing and assembly

The BAC DNA was sequenced on an *Illumina* platform by producing 100bp paired-end reads. The sequencing library was prepared by the sequencing service team at TGAC (Norwich), following *Illumina*'s protocol for Paired-End DNA sample preparation. The sequencing reads were filtered based on quality as detailed in section 2.7.2.1 and assembled into three contigs (60kb, 25kb and 7.5kb long) by Annabel Whibley (Coen Lab, JIC).

The assembly was performed using two software tools in parallel: Velvet 1.2.03 (Zerbino & Birney 2008) and ABySS 1.3.7 (Simpson *et al.* 2009). In both cases, the option for “k-mer length” was set to 59. The two assemblies were compared by manual alignment and a consensus assembly produced in Geneious (Biomatters Limited).

One of these contigs was used to bridge a gap between three scaffolds in the *A. majus* reference sequence (version 1.0). The assembled BAC contig and the three reference genome scaffolds were manually aligned and merged into a consensus sequence using BioEdit (version 7.2.5). This consensus scaffold (*ROS* scaffold) replaced the following scaffolds in the *A. majus* reference genome (version 1.0): scaffold117, scaffold678 and C8923637.

2.7 High-throughput sequencing techniques

High-throughput sequencing was used throughout this work for the sequencing of a BAC plasmid containing the *ROS1* gene (section 2.6), whole-genome sequencing of pooled DNA from wild *A. majus* populations (chapter 3) and whole-transcriptome sequencing of RNA extracted from flower buds (chapter 5). In all cases, sequencing was performed on an *Illumina HiSeq2500* machine at TGAC (Norwich).

Table 2.2 – Summary statistics of *A. majus* reference genome.
Version 1.0 of the genome was used in this work. "N50" is defined as the length of which all scaffolds of that length or longer add up at least half the total of the lengths of all scaffolds.

Total scaffold size	500 Mb
N50 scaffold size	585,952 bp
Scaffold number	87,577
> 100 bp	85,573
> 2 kb	2,510
GC content	35.2%

The reference genome used to map the *Illumina* sequence data was of an inbred line of *A. majus* (*JI7* stock line). This genome is currently being assembled and annotated at BGI (Beijing, China) as part of the "1000 plant genomes" project and is not yet publicly available (it was made available to the *Antirrhinum* research community by Yongbiao Xue). The genome version used in this work (version 1.0) consists of 87,577 scaffolds of varying size (Table 2.2) that have been grouped into 8 linkage groups (corresponding to the 8 *Antirrhinum* chromosomes), using a mapping population of 48 recombinant inbred lines derived from an *A. majus* x *A. charidemi* cross produced in the Coen lab (Norwich). For this thesis, the reference genome was used as is, except for scaffolds linked to the *ROS1* gene, which were modified as explained in section 2.6.3.

2.7.1 Preparation of material

DNA and RNA were extracted as detailed in sections 2.3.2, 2.3.5 and 2.6.2.

The sequencing libraries were constructed by the service provided at TGAC (Norwich), which involved the quality control of the DNA/RNA samples, and the construction and quality control of the sequencing libraries (following the relevant *Illumina* library preparation protocols for each case).

DNA was sequenced using 100bp paired-end reads. mRNA was sequenced using 50bp single-end reads.

2.7.2 Bioinformatics analysis

All programs were run with default options, unless indicated otherwise.

2.7.2.1 Read quality filtering

Reads in FASTQ format were filtered based on their quality (as determined by the standard *Illumina* quality phred score in Sanger format), by using the software *fastq-mcf* (Aronesty 2011, available at http://code.google.com/p/ea-utils; accessed March 2014). Reads were trimmed after a position with a quality score below 20 was found (option "-q 20"). After trimming, only reads with a minimum length of 30 bp (option "-l 30") and minimum mean quality score of 20 (option "--qual-mean 20") were kept. Reads with ambiguous bases were discarded (option "--max-ns 0"). All sequencing cycles were retained (option "-k 0"); the adapter clipping option was turned off (specified "n/a" in command). In summary, the following non-default command line options were used:

```
-l 30 -k 0 -q 20 --qual-mean 20 --max-ns 0 n/a
```

2.7.2.2 mRNA sequencing pipeline

This pipeline was used for analysing the sequenced mRNA samples detailed in chapter 5. There were 11 samples, each deriving from RNA extracted from the corolla of *A. majus* flower buds as detailed in section 2.4.

The filtered FASTQ reads (50 bp single end) from each sample were mapped to the *A. majus* reference genome (version 1.0) with the modified *ROS* scaffold (see section 2.6) using the software *tophat v. 2.0.4* (Trapnell *et al.* 2009). During alignment, mapping seeds were allowed to have 1 mismatch (option "-b2-N 1") and the final read alignments were allowed to have up to 4 mismatches (option "-N 4").

The maximum intron size was set to 30 kb (option "-I 30000"). In summary, the following non-default options were used:

```
-N 4 -I 30000 --b2-N 1
```

This tool outputs a file in *.bam* format (11 files, one for each sample).

The mapped reads were assembled into transcripts by using the *cufflinks* tool from the *cufflinks v. 2.1.1* software (Roberts *et al.* 2011). The option to correct for reads mapping to multiple locations was turned on (option "-u") and transcript expression was normalized by the upper quartile of the number of fragments mapping to each gene, to avoid biases due to unusually highly expressed transcripts (option "–N"). Intron size was kept to a maximum of 30 kb (option "-I 30000"). In summary, the following non-default options were used:

```
-I 30000 -u -N
```

Because a transcript assembly was obtained for each of the 11 samples separately, these assemblies were combined into a single file using the *cuffmerge* tool from the *cufflinks* package, using default options. This provided a consensus transcript assembly (*.gtf* format) that was used for calculating each transcripts' expression value in the next step of the pipeline.

Normalized expression values (*RPKM*; detailed in section 5.1) were calculated for each transcript using the *cuffdiff* tool in the *cufflinks* package. The following non-default options (detailed above for *cufflinks*) were used: `–N –u`. This tool outputs several files, but for this work, only the one named "genes.fpkm_tracking" was used. This file contains normalized expression values (*RPKM*) for each assembled gene.

The output from *cuffdiff* was analysed in the statistical package *R* (version 3.0.1) to perform the analysis discussed in chapter 5.

2.7.2.3 Whole-genome sequencing pipeline

This pipeline was used for analysing the samples detailed in chapter 3. There were 6 samples, each consisting of a DNA pool from 50 or 52 individuals collected from the *pseudomajus* x *striatum* hybrid zone.

The filtered FASTQ reads (100 bp paired-end) from each sample were mapped to the *A. majus* reference genome (version 1.0) with the modified *ROS* scaffold (see section 2.6) using the software *stampy v. 1.0.20* (Lunter & Goodson 2011). Because the sequenced samples were expected to have some divergence in relation to the reference, the expected divergence option was set higher than the default, to 5% (option "--substitutionrate=0.05"). This tool outputs a file in *.sam* format (6 files, one for each sample).

After mapping, read duplicates were removed from each *.sam* file using the *MarkDuplicates* tool included in the *Picard v. 1.107* software (http://picard.sourceforge.net; accessed March 2014). An indexed BAM file (*.bai*) was created in this step using the following non-default option: `CREATE_INDEX=true`. The output consisted of *.bam* and *.bai* files for each of the 6 samples.

For estimating a measure of within-population diversity (π) the software package *popoolation* (Kofler *et al.* 2011a) was used.

First, each of the 6 *.sam* files was converted to *pileup* format using *samtools v. 0.1.18* (Li *et al.* 2009). The minimum mapping quality score for a read to be considered was set to 20 (option "-q 20"); the minimum quality score for a base to be considered was set to 20 (option "-Q 20"); the read depth was kept in the output (option "-D") and anomalous read pairs kept (option "-A"). In summary, the following non-default options for the *samtools mpileup* tool were used: `-q 20 -Q 20 -BDA`. This resulted in 6 files (one per sample) in the *pileup* format.

Second, the *Variance-sliding.pl* script from the *popoolation* package was used to calculate window-averaged measures of π. The following criteria were used for each position to be considered for analysis: the depth of coverage had to be between 10x - 200x (options "--min-coverage 10 --max-coverage 200"); a minimum of 2 read counts was necessary for a SNP to be considered (option "--min-count 2"); and the minimum base quality score had to be 20 (option "--min-qual 20"). In summary, the following options were used for all samples:

```
--measure pi --pool-size 100 --fastq-type sanger --min-count 2
--min-coverage 10 --max-coverage 200 --min-covered-fraction 0
--min-qual 20 --window-size 10000 --step-size 5000.
```

Some options were variable: `--pool-size` had the value 100 (for pools composed of 50 plants) or 104 (for the pool composed of 52 plants); `--window-size` and `--step-size` also varied depending on the averaging window size and the step size between each window that was desired.

For estimating a measure of between-population divergence (*Fst*) the software package *popoolation2* (Kofler *et al.* 2011b) was used.

First, the *.sam* files from each of the 6 mapped samples was converted to a single file in the *mpileup* format using *samtools v. 0.1.18* (Li *et al.* 2009). The following options (detailed above) for the *samtools mpileup* tool were used:

```
-q 20 -Q 20 -BDA.
```

The output was a single file (combining all 6 samples) in the *mpileup* format.

Second, the *mpileup* file was converted to a format required by the *popoolation2* package, using its tool *mpileup2sync.jar*, with default options.

Third, the converted file was used to calculate window-averaged *Fst* using the *popoolation2* script *fst-sliding.pl*, with the following options (detailed above):

```
--min-count 2 --min-coverage 10 --max-coverage 200 --min-covered-fraction 0 --pool-size 104:100:100:100:100:100.
```

As before, the options `--window-size` and `--step-size` varied depending on the desired averaging window size and step size between each window.

For obtaining a read count of each SNP in the genome, the *popoolation2* script *snp-frequency-diff.pl* was used, with the following options (detailed above):

```
--min-count 2 --min-coverage 10 --max-coverage 200.
```

The outputs from the *popoolation* and *popoolation2* scripts were analysed in the statistical package *R* (version 3.0.1) to perform the analysis discussed in chapter 3.

2.8 Fitting of clines in the hybrid zone

In Figure 6.5B the haplotype frequency changes for markers along the *pseudomajus* x *striatum* hybrid zone transect were fitted with two curves: a 4-parameter sigmoidal function and a Gaussian function. In both cases, the data were fitted to each model using the *nls* function in the statistical package *R* (version 3.0.1).

The 4-parameter sigmoidal function used for *A. m. pseudomajus* and *A. m. striatum* *ROS1 EL-MYB* haplotypes was:

$$p = \frac{b - a}{1 + e^{\frac{4\times(c-x)}{w}}} + a$$

Where a and b are the left and right asymptotes of the curve, respectively. x is the geographic position of a sample along the transect; c is the centre of the cline; w is the width of the cline. Except for x, which is known from the data, the other four parameters (two asymptotes, cline width and cline centre) were estimated by the *nls* fitting procedure.

The Gaussian function used for the recombinant *ROS1 EL-MYB* haplotypes was:

$$p = a \times e^{-\frac{(x-c)^2}{2\times w^2}}$$

Where a is the height of the curve peak; x is the geographic position of a sample along the transect; c is the centre of the bell curve; and w is related to the width of the bell curve. Except for x, which is known from the data, the other three parameters (height, centre and width of the curve) were estimated by the *nls* fitting procedure.

These model fittings will be improved in the future by using a maximum likelihood approach and trying out different models (Barton & Gale 1993).

3 Genomic divergence around ROSEA locus

Work by Whibley et al. (2006) revealed that the *ROS1* gene is divergent between the two populations forming the *pseudomajus* x *striatum* hybrid zone. The main result supporting this divergence was the description of a sharp allelic cline for this locus across the hybrid zone transect. This contrasts with other linked markers that do not show such a pattern, namely the *DICH* and *PAL* loci, classically mapped as 9 cM and 16 cM away from *ROS*, respectively [Stubbe (1966) cited in Whibley (2004)]. In this chapter, I explore how extensive the divergence is around the *ROS1* gene, using genome-wide sequence data. Several pools of randomly sampled plants were collected at different distances from the centre of the flower colour cline (three "magenta" pools and three "yellow" pools; Table 3.1). The pools were composed of 50-52 plants each (i.e. equivalent to 100-104 haploid genomes) and were whole-genome sequenced (methods 2.7). I used these data to calculate measures of nucleotide diversity (π) and divergence (*Fst*) across the genome. I found that a narrow (< 1cM) peak of *Fst* precicely coincides with the colour gene *ROS1*. Concordant with the colour and allelic clines, this pattern was observed only in comparisons between samples from oposite sides of the cline (i.e. magenta-yellow comparisons), but not between those from the same side (i.e. magenta-magenta and yellow-yellow comparisons). Other linked narrow peaks of *Fst* occur downstream of the *ROS1* peak, suggesting that other loci in the region might be under selection. This heterogeneous profile of divergence fits with the current view that the genomic landscape of divergence between populations sharing a recent ancestor and/or undergoing gene-flow is composed of "islands of divergence" (presumably containing loci under selection) on a "sea" of lower divergence (presumably containing neutral variation). The peaks of *Fst* in the *ROS* region have slightly reduced intra-population diversity (π), which might suggest past or current selective events that established different alleles in the two *A. majus* subspecies (although other alternative explanations are considered). This chapter demonstrates how sampling genomes across a phenotypic cline can precisely pinpoint putative loci important for reproductive barriers.

Table 3.1 - Summary of pooled samples used for WGS

ID	Distance from centre [a] (Km)	Pool size[b]	Sequencing depth median *D* [c]	% Genome sequenced [d]	Mean SNP density [d] (per kb)	Total no. SNPs [d] (10e6)
YP4	-12.9	52	44	62	44	27.7
YP2	-1.6	50	25	73	31	19.6
YP1	-1.8	50	23	72	30	19.0
MP2	0.7	50	23	72	30	19.9
MP4	1.4	50	26	74	31	20.0
MP11	8	50	34	57	36	21.8

[a] Euclidian distance from the canonical centre of the hybrid zone flower colour cline. Negative and positive distances correspond to pools on the West and East of the centre, respectively.

[b] Number of diploid individuals included in the pool.

[c] Sequencing depth is given as median read count in each position of the genome. This excludes non-sequenced bases mapped to the *A. majus* reference genome

[d] Considering mapping quality ≥ 20; read quality ≥ 20; sequencing depth (D) between 10-200

3.1 Introduction: Measures of population diversity and divergence

Two main population genetics statistics will be used throughout this chapter: π, which is a measure of nucleotide diversity within a particular population and *Fst*, which is a measure of divergence between two populations. The two statistics are related and calculating one allows calculation of the other. I will consider these two measures in the context of a single nucleotide polymorphism (SNP) with only two alleles.

π is defined as the average number of pairwise differences between sequences in a sample (Hedrick 2011), which is equivalent to the expected heterozygosity (*H*) in the population. For convenience, I will focus on expected heterozygosity, which involves a simple and intuitive calculation. However, π is related to it, since they are both measures of the allelic diversity in a population. The expected heterozygosity is calculated as the chance of drawing two different alleles sampled from a population, which is given by 1 minus the probability of drawing two equal alleles:

$$H = 1 - (p^2 + q^2)$$

Where, p and q are the frequencies of two alleles (e.g. in a SNP). This measure varies between 0 and 0.5. In one extreme, if a SNP is not polymorphic within a population, then there is no genetic variation and $H = 0$. On the other extreme, if both alleles are at 50% frequency in the population, the genetic diversity is maximized and $H = 0.5$. Note that the observed heterozygosity in a population (i.e. frequency of heterozygotes) might be different from the expected heterozygosity. Consider the following two extremes for a diploid population: if there are only heterozygous individuals, the observed heterozygosity is 1, but the expected heterozygosity is 0.5 (because each allele is at 50% frequency in the population); conversely, in a population composed of only homozygotes, but with each of two homozygote classes at 50%, the observed heterozygosity is 0, but the expected heterozygosity is still 0.5. Thus, estimating expected heterozygosity is informative about how variable a region is at the nucleotide level in a population, but it is uninformative about how each allele is coupled in the individuals composing that population.

Measuring heterozygosity (or π) is useful to detect events associated with reductions in nucleotide diversity. For example, if a population went through a recent bottleneck, the nucleotide diversity in that population will come from the few individuals that survived, whereas variation from other non-surviving individuals is lost. A reduction of nucleotide diversity can also be seen around a recently positively selected locus, a phenomenon known as a selective sweep. This is because, as a selected allele increases in frequency in the population, neutral polymorphisms linked to it will also increase in frequency (genetic hitchhiking), resulting in a highly prevalent haplotype in the population, whereas non-selected haplotypes disappear from the population. The converse phenomenon might also occur: if there is negative selection against a deleterious mutation, variation linked to that mutation will be removed by selection, reducing diversity in and around the selected locus (background selection). These signatures of selection will disappear over time, as new mutations accumulate restoring diversity in the region (although this might take a long time, since the mutation rate in eukaryotes is relatively low, on the order of 10^{-9} mutations per generation per site; Lynch 2010).

Whereas heterozygosity deals with the diversity found in one population, *Fst* measures population divergence by comparing the nucleotide diversity between two populations with that within each population. It is calculated as:

$$Fst = \frac{H_T - H_W}{H_T}$$

Where H_T is the expected heterozygosity of the entire meta-population (i.e., considering both populations together) and H_W is the average expected heterozygosity within each population ($H_W = \frac{H_1+H_2}{2}$). *Fst* varies between 0 and 1, with higher values indicating a greater divergence (higher values mean that the variation between the populations is greater than the variation within each population). For example, if two populations are fixed for a different allele, each of them will have $H = 0$, and therefore $H_W = 0$. This will result in $Fst = \frac{H_T - 0}{H_T} = 1$. Conversely, if both populations have the same allele frequencies, then $H_T = H_W$ and $Fst = 0$ (more examples in Figure 3.1).

Fst is thus useful to find loci that are differentiated between populations, suggesting divergent selection acting on them (although other selective or even non-selective scenarios might also elevate *Fst*, discussed later). Even if the average π around a selected locus is not reduced in either population, as could be expected if one of them went through a recent selective sweep, the average *Fst* may still be significantly greater than 0 (Figure 3.1).

The data presented in this chapter use measures of π and *Fst* implemented in the *popoolation* and *popoolation2* packages (Kofler *et al.* 2011b; a). These differ from the formulas presented above in that they are corrected for dealing with whole-genome sequence data from pooled DNA (i.e. DNA extracted from a pool of sampled individuals). However, this technical detail does not change the interpretation of their values explained here.

SNP	**1**	**2**	**3**	**4**	**5**	**Average SNPs 1-5**
Population 1	A	A	A	T	A	
	A	A	A	T	A	
	T	A	A	T	A	
	T	A	T	T	A	
	T	T	T	T	A	
	T	T	T	T	A	
Population 2	A	A	T	A	T	
	A	A	T	A	T	
	T	T	T	A	T	
	T	T	T	T	T	
	T	T	T	T	T	
	T	T	T	T	T	
H_1	**0.44**	**0.44**	**0.5**	**0**	**0**	**0.376**
H_1	**0.44**	**0.44**	**0**	**0.5**	**0**	**0.376**
H_W	0.44	0.44	0.25	0.25	0	
H_T	0.44	0.5	0.375	0.375	0.5	
Fst	**0**	**0.11**	**0.33**	**0.33**	**1**	**0.354**

Figure 3.1 – Example calculations of heterozygosity (*H*) and divergence (*Fst*).

Examples for five bi-allelic loci (represented as SNPs) in two populations. The examples show, from left to right: a polymorphism occurring at equal frequencies in both populations (SNP 1); a polymorphism occurring at different frequencies in each population, even though each population has the same heterozygosity (SNP 2); a polymorphism occurring in only one of the populations (SNPs 3 and 4); a polymorphism fixed between populations (SNP 5). The average *Fst* and heterozygosity of each population across the 5 SNPs is shown. Statistics are calculated as explained in the text.

3.2 Results

3.2.1 Criteria for analysis of genome sequence data

The use of high-throughput sequencing of DNA pools from natural populations is an attractive approach for accessing the allelic variation in a population sample, as it avoids the cost of sequencing many individual samples (Futschik & Schlötterer 2010; Kofler *et al.* 2011a). However, the accuracy of allele frequency estimates in pooled samples is limited by certain characteristics of whole-genome datasets. For this reason, it is important to define those features and see how they affect downstream analysis.

The raw genome sequencing data (e.g., generated by *Illumina* platforms, as in this study) consist of several millions of short sequences (*reads*), usually between 50-150bp long. A typical analysis pipeline for this kind of data consists of (i) filtering the reads based on their sequence quality (quality filtering), (ii) matching each read to their corresponding position in a reference genome (mapping) and (iii) identifying SNPs in relation to the reference genome (see section 2.7.2 in methods for the pipeline used in this work). The reads are randomly obtained from the DNA sample being sequenced and, due to the large sequence output, each particular base of that sample is sequenced several times. For example, if a sequencing run outputs a total of 2000 Mb and the DNA sample being sequenced is 100 Mb long, then each position in that sample should be sequenced, on average, 2000/100 = 20 times. This concept of how many times a particular base is sequenced is often referred to as "depth of coverage" or "sequencing depth", which I will simply call *D* (Figure 3.2). Because the sequencing procedure is random, there is a distribution around the expected mean of *D* (i.e. most bases of that sample will be sequenced more or less times than the average).

Estimating allele frequencies from pools takes advantage of *D* in next-generation sequence data. The main assumption is that, for each position in the genome, *D* is the size of a sample of the chromosome copies present in the DNA pool. Therefore, the proportion of a nucleotide in *D* is expected to provide an estimate of its frequency in the pool of individuals. For example, if a base was sequenced 40 times, i.e. *D* = 40, and of those bases 10 were the nucleotide A and the remaining 30 the nucleotide T, then the proportion of "A" is *10/40 = 0.25* and that of "T" is *30/40 = 0.75*. These should be proportional to the actual frequency of each allele ("A" and "T" in the example) in the pool of sequenced individuals; in this population it would mean that the "A" allele is the less common one. Sampling biases have to be taken into account (for example as *D* gets lower, the probability of missing an allele increases), and studies investigating these issues demonstrate improved estimates are obtained with (i) higher sequencing depth (*D)* and (ii) higher numbers of individuals included in each pool (Futschik & Schlötterer 2010; Ferretti *et al.* 2013; Anderson *et al.* 2014). *D* is important because, as mentioned, a higher *D* is equivalent to a larger sample size of the chromosome copies contained in the pool. Higher numbers of individuals included in the pool effectively "dilutes" each chromosomal copy more, reducing the probability that a particular copy over-dominates the sample (due to PCR sequencing bias), leading to distorted frequency estimates. A higher *D* is always desirable, but there is a trade-off between the idealized dataset and the cost of obtaining it.

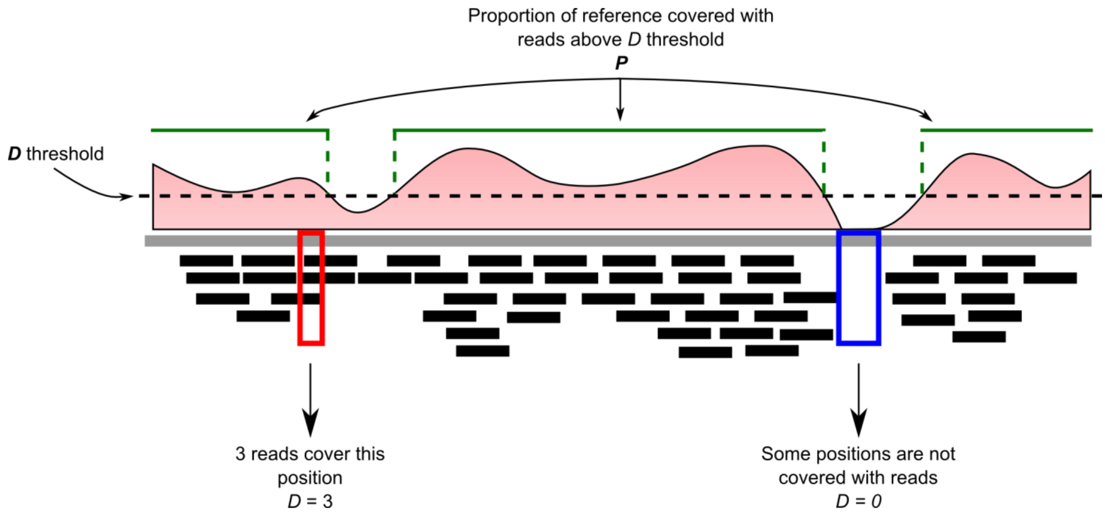

Proportion of reference covered with reads above *D* threshold
P
D threshold
3 reads cover this position
D = 3
Some positions are not covered with reads
D = 0

Figure 3.2 – Schematic view of key features of whole genome datasets.
The scheme represents a portion of a genome (grey bar) to which several reads (black lines) were mapped. The sequencing depth, *D*, is the number of reads overlapping a position of the genome. Because the sequencing procedure is random, *D* is variable across the genome (shown as a histogram in pink), and some regions may even be missed (purple box). The proportion of the genome (or a window in the genome) which satisfies a minimum *D* threshold is denoted by *P* (green lines above histogram).

As mentioned, I will use *Fst* as the main measure of divergence between the sequenced DNA pools. Because *Fst* requires estimating allele frequencies (section 3.1) and these, in turn, are affected by the sequencing depth *D*, *Fst* estimates from pooled data can be noisy. One approach to overcome this noise is to assume that physically linked regions in the genome have similar patterns of diversity and therefore statistics can be averaged over windows across the genome (the size of those windows can be variable, in this study I will use 10kb windows). This reduces the contribution of spurious noise from each individual SNP, allowing the identification of consistently highly divergent regions.

The approach of using window-averaged statistics requires the introduction of another important concept in genomic datasets: the fraction of the window that is sequenced (sometimes referred to as “physical coverage”). Some regions of the genome may not be covered with reads, either by chance (if no reads were recovered overlapping a particular base), due to differences between the individuals being sequenced and the sequence being used as a reference (in this study the reference genome is from an inbred line of *Antirrhinum majus*, which may differ from *A. m. pseudomajus* and *A. m. striatum*) or due to repetitive regions in the genome, where mapping is ambiguous (this reduces the mapping quality score of the reads). Furthermore, a particular base may have such low *D* that it is excluded from the analysis (usually a threshold for *D* is defined, as discussed below). I will denote the proportion of bases in a window with a minimum sequencing depth *D* as *P* (Figure 3.2). *P* is important, as it is also related to the noise in the dataset. For example, a 10kb window with *P = 1%* will only contain *10000 x 0.01 = 100* sequenced bases. If, from these, only one base is polymorphic, the *Fst* for that window is calculated from a single SNP, which defeats the purpose of having window-averaged estimates. Considering the importance of *D* and *P* in estimating window-averaged *Fst* from pooled datasets, I explored how varying both parameters affects the amount of usable data and the quality of the data. I focus on two of the sequenced pools (pools YP1 and MP4 in Table 3.1), but the results presented here are similar for other comparisons.

Firstly, I explored how many windows, from the total across the *A. majus* reference genome (129,387 sliding-windows of 10kb length and a step size of 5kb), are left when setting different minimum *D* and *P* thresholds. Increasing the *D* threshold has a considerable impact in the fraction of usable windows (Figure 3.3A). For example, if *D ≥ 30x* and *P ≥ 1%* only 40% of the total number of 10kb windows is usable, and this effect is worsened as the *P* threshold increases (Figure 3.3A). This is possibly related to the samples used having a median *D* between 23 and 44 (Table 3.1). Indeed, the window loss is less severe for thresholds of *D ≥ 10x* or even *D ≥ 20x* (Figure 3.3A). Therefore, the possibility of setting a threshold for *D* ≥ 30x was discarded as it eliminates too large a proportion of the dataset.

Given these results, I considered using either a threshold of *D ≥ 20x* (which should improve allele frequency estimates) or *D ≥ 10x* (which maximizes the number of usable windows). It is difficult to define a canonical set of parameters that optimize the balance between having a sufficient amount of data and the accuracy of the data. For analysis of divergence, like *Fst*, one is mainly interested in regions of high divergence, which are characterised by polymorphisms where each population is fixed for a different allele (for example, SNP 5 in Figure 3.1). Therefore it is useful to calculate the probability of getting a fixed difference by chance. Let us assume two alleles, A and T, at a frequency of 0.5 in each of two samples (i.e., there is no difference between allele frequencies within and between samples, *Fst = 0*). The probability of obtaining a fixed difference by chance is given by the probability (*P*) that in each sample only one of the alleles was sequenced, that is:

$$P = \ p(sample1 = A) \times p(sample2 = T) + \ p(sample1 = T) \times p(sample2 = A)$$

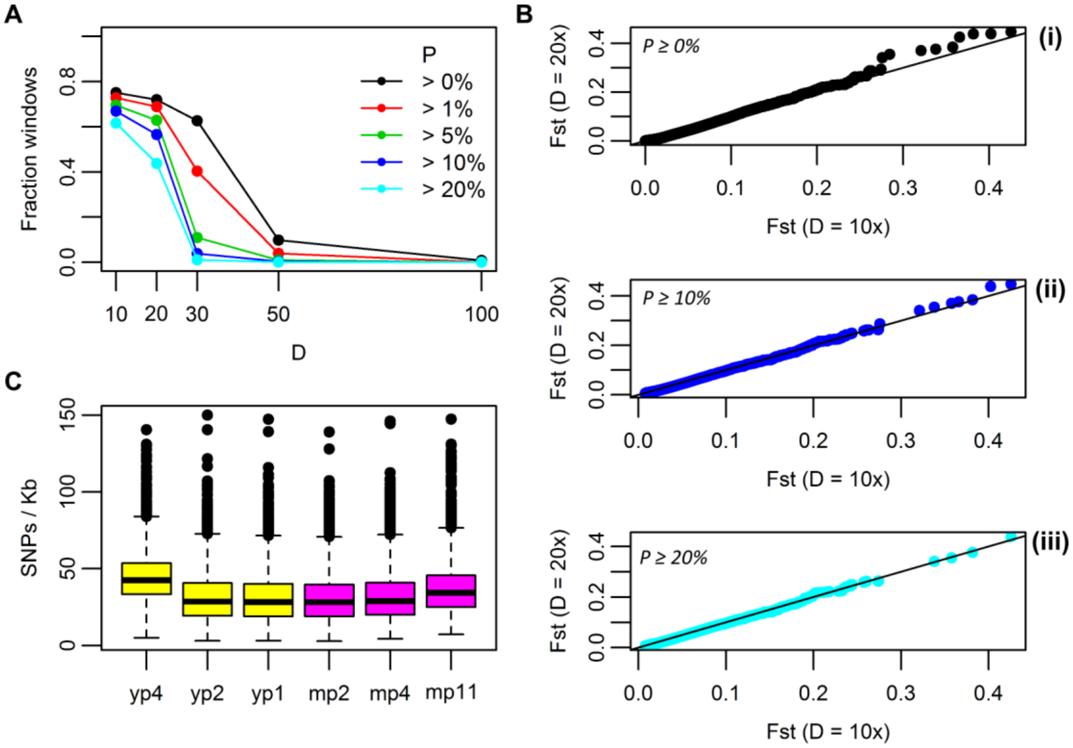


Figure 3.3 - Exploring different sequencing criteria for analysis of whole-genome-sequence datasets, using the empirical data from sequenced DNA pools.

A) Fraction of 10kb genomic windows fulfilling certain *D* and *P* thresholds. The y-axis plots the fraction of windows in relation to the total in the *A. majus* reference genome. The criteria used are: (i) every site within a window has *D* as indicated by the x-axis; and (ii) the window is covered to a minimum *P* indicated by the coloured lines.

B) Quantile-Quantile plots (QQ-plots) for 10kb-window-averaged Fst distributions using sites with *D* ≥ 10x (x-axis) and *D* ≥ 20x (y-axis). The QQ-plot compares two distributions by looking at the quantile values of each. If the distributions are similar, the points fall on the y=x line (black line). The three plots (i to iii) show the QQ-plot using different minimum *P* thresholds.

C) Distributions of number of SNPs per kb for each pool. These distributions are for SNPs with *D* ≥ 10x. The points above the boxplot wiskers are values above 1.5x the third quartile of each distribution.

Plots in **A)** and **B)** are for 10kb windows, but the same patterns are true for 5kb and 50kb windows.

Assuming *D = 10x* (this is our sample size), then the probability of sampling just one of the alleles and missing the other is 0.5^{10}. Therefore, the above equation becomes:

$$0.5^{10} \times 0.5^{10} + 0.5^{10} \times 0.5^{10} \sim \mathbf{10^{-6}}$$

The effect of false fixed polymorphisms is further ameliorated when using window-averaged statistics. For example, if there are two SNPs in a window, the probability that both of them are fixed by chance is 10^{-12}, for three SNPs 10^{-18}, etc... These probabilities are already quite low, and increasing *D = 20x* simply reduces them further (doubling the exponent of each probability exemplified for *D = 10x*).

To empirically assess if using *D ≥ 10x* or *D ≥ 20x* produces significantly different estimates of *Fst*, I compared whole-genome *Fst* distributions under the two conditions. To do this comparison, I used a quantile-quantile plot (QQ-plot), which is a graphical method for comparing distributions against each other. If two distributions are exactly the same, their quantiles match perfectly and this is seen on a QQ-plot as points lying along on the *y = x* line. The *Fst* distributions with *D ≥ 10x* and *D ≥ 20x* are overall well correlated, even when the window coverage is ignored, that is, *P ≥ 0%* (Figure 3.3Bi). However, in the upper tail of the distribution, using *D ≥ 10x* results in a lower estimate of *Fst* compared with *D ≥ 20x* (see points not overlaying *y = x* line in Figure 3.3Bi). This discrepancy is reduced when *P ≥ 10%* and virtually disappears when *P ≥ 20%*. Indeed, the strength of the *Fst* correlation between *D ≥ 10x* and *D ≥ 20x* increases as more sites are included in the windows (linear correlation r^2 = 0.61, 0.76 and 0.84 for *P ≥ 0%*, *P* ≥ 10% and *P* ≥ 20%, respectively).

Considering this exploratory analysis, I settled on the following criteria for the remainder of my window-based analysis: *D ≥ 10x* for each site and *P ≥ 20%* for each window. These criteria allow the inclusion of ~61% of all 10kb windows in the genome (79,807 windows of size 10kb, sliding across the genome in 5kb steps). For individual SNP analysis (rather than window averages) the higher threshold of *D ≥ 20x* was used.

For *D ≥ 10x* there are, on average, 30-44 SNPs/Kb (Table 3.1; Figure 3.3C). Therefore, a 10kb window with *P = 20%* (the minimum threshold), will have 2kb of sequenced bases and, on average, 60-88 SNPs. In fact, most windows are likely to have more SNPs than this, since most of them have *P ≥ 20%* (the median window *P* is 63%). Finally, since several pools were sequenced, divergence patterns can be further confirmed by looking at several *Fst* pairwise comparisons between them, as shown in the next section.

3.2.2 Divergence around ROS locus

To explore the divergence across the *ROS* locus I used the genome-wide data to compute window-averaged *Fst* as implemented in the *popoolation2* package (Kofler *et al.* 2011b) and using the criteria defined in the previous section (*D ≥ 10x* and *P ≥ 20%*) (section 2.7 in methods). I identified three scaffolds in the *A. majus* reference genome containing the *ROS1* gene and some of its known flanking sequence. To close the gaps between those scaffolds, a previously identified BAC clone containing *ROS1* was fully sequenced and assembled into three contigs (section 2.6 in methods). I used this assembly to bridge the gaps between the genome scaffolds and constructed a consensus assembly which now consists of a single scaffold ~927kb long, containing *ROS1* in position ~541kb. For the remaining of this chapter I will focus my attention on this scaffold (named "*ROS* scaffold"), which is located in Linkage Group 6 (LG6).

I firstly compared two pools from opposite sides of the flower colour cline (YP1 and MP4; Table 3.1). One pool is located 1.8km away from the centre of the hybrid zone towards the *A. m. striatum* side (the yellow pool, YP1) and the other pool is located 1.4km towards the *A. m. pseudomajus* side (the magenta pool, MP4). These two pools are close to the hybrid zone centre, but on the edges of the allelic cline defined by *ROS1*. Therefore, they represent two samples where the divergence linked to *ROS1* can be investigated. The extent of divergence around *ROS1* is evident in the variation of *Fst* along LG6 (Figure 3.4A). *Fst* along this linkage group is variable, but the *ROS* scaffold contains clear outlier windows, reaching a value of *Fst =* 0.43 (this is the maximum value across the whole genome between these two samples; Figure 3.4B). The scaffolds containing the genes *DICH* and *PAL* show a *Fst* signal that is comparable to the genome-wide average (blue and cyan arrows in Figure 3.4A), agreeing with previous work that showed these genes do not to exhibit an allelic cline across the hybrid zone (Whibley *et al.* 2006).

In more detail, *ROS1* perfectly co-localizes with a main peak of *Fst* (pink arrow in Figure 3.4C). This result strongly suggests that the *Fst* peak is not due to random processes (e.g. sampling bias, or incomplete mixing of neutral variation across the hybrid zone cline), but rather due to divergent selection acting on flower colour. The profile of divergence across this region is quite heterogeneous with at least two other prominent peaks occurring downstream of *ROS1* and interspersed by near-baseline levels (arrows in Figure 3.4C).

Fst is a summarised measure of nucleotide divergence, which may include several kinds of SNPs. Therefore, I qualitatively classified SNPs in three ways: a “fixed” SNP occurs when each population is fixed for a different allele (e.g. SNP 5 in Figure 3.1); a “shared” SNP occurs when both populations are segregating for two alleles (e.g. SNP 1-2 in Figure 3.1); and a “private” SNP occurs when one population is segregating for two alleles but the other population is fixed for one of the alleles (e.g. SNPs 3-4 in Figure 3.1). “Fixed” SNPs point to regions which are not exchanged between populations, as would be expected for loci under divergent selection. Conversely, “shared” SNPs arise either through gene flow between populations, as would be expected for neutral loci, or can be due to ancestrally shared variation.

I considered a SNP to be “fixed” if the allele frequency difference between the pools was ≥ 90%, and “shared” if the minor allele frequency was ≥ 20% in both samples. To visualize the occurrence of these two kinds of SNPs across the *ROS* scaffold, I coloured each class in Figure 3.4B. Co-localizing with the *Fst* peaks there is an excess of “fixed” polymorphisms and a shortage of “shared” ones. However, between these regions, not only can “fixed” SNPs be absent, but “shared” SNPs occur at frequencies comparable to those observed elsewhere in the genome.

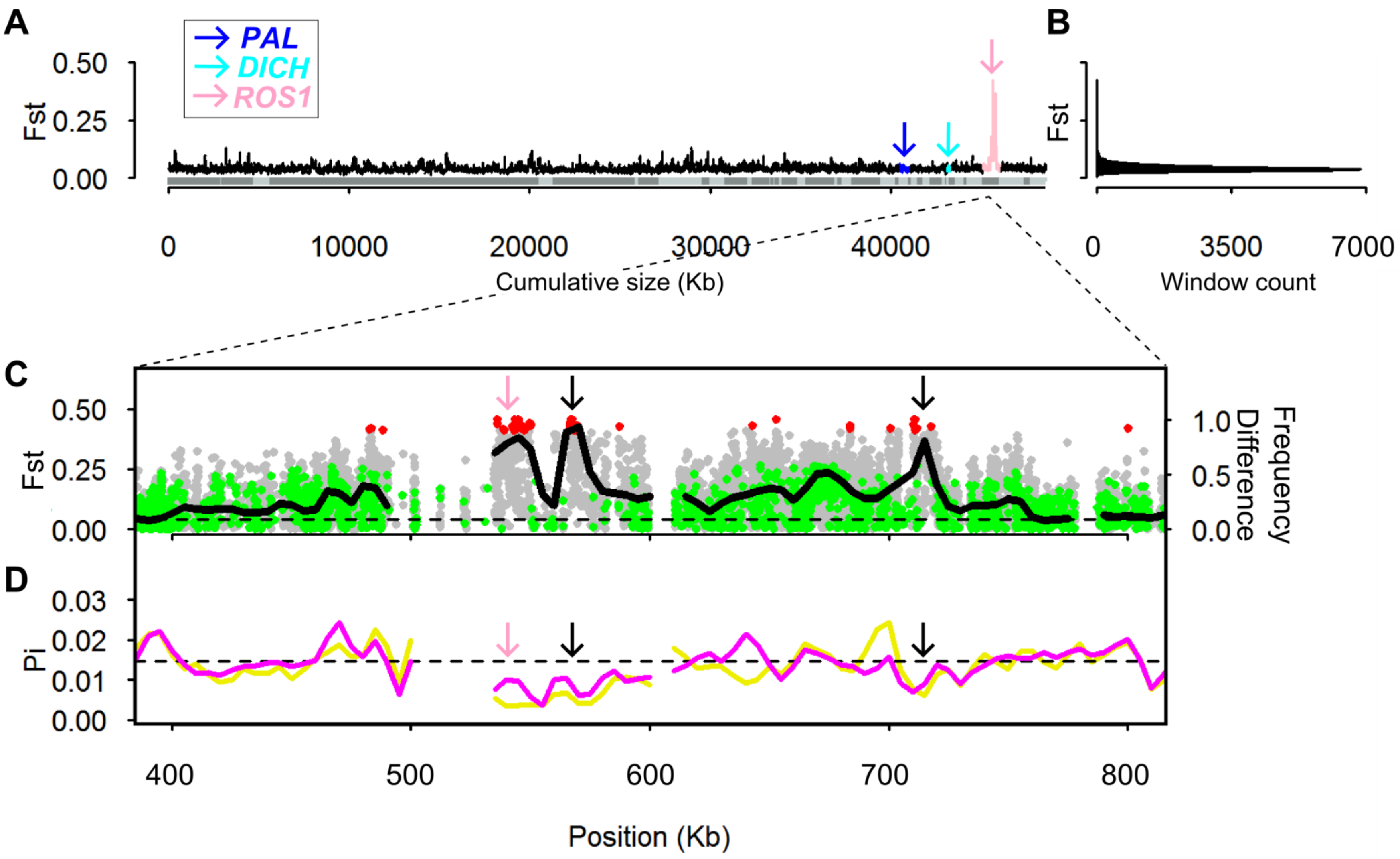

A
PAL
DICH
ROS1
0.50
Fst 0.25
0.00
0
10000
20000
30000
40000
Cumulative size (Kb)
B
Fst
0
3500
7000
Window count
C
0.50
Fst 0.25
0.00
1.0
0.5
0.0
Frequency Difference
D
0.03
0.02
Pi
0.01
0.00
400
500
600
700
800
Position (Kb)

Figure 3.4 (previous page) – Divergence around *ROS1* locus between MP4 and YP1 pools.
A) *Fst* across *A. majus* linkage group 6. Map intervals (blocks where no recombination between scaffolds is available) are indicated as two intercalated shades of grey in the x-axis. Scaffolds within each map interval are ordered randomly. The values of *Fst* are 10kb window averages. The scaffolds containing the genes *PAL*, *DICH* and *ROS1* are coloured as in the legend. The position given in the x-axis is the cumulative size of the scaffolds in Kb.
B) Genome-wide *Fst* distribution. Notice the long narrow tail of higher values, of which the *ROS1* scaffold contains the maximum for the comparison between these two pools.
C) A zoomed-in view of **A)**, showing the variation around the *ROS1* region. The *Fst* (line) and allele frequency difference (points) between the two pools is shown. The points are for individual SNPs with *D ≥ 20x*; "fixed" and "shared" SNPs are coloured in red and green, respectively. The horizontal dashed line is the genome-wide *Fst* median. Arrows point to three prominent *Fst* peaks (pink arrow locates *ROS1* gene). The position given in the x-axis corresponds to the position within the *ROS* scaffold. Gaps in the plot line are windows where the coverage was below the set threshold of *P* = 20%.
D) Nucleotide diversity (π, averaged over 10kb windows) across the region shown in A). π is shown for both pools being compared (MP4 – magenta line; YP1 – yellow line). The horizontal dashed line is the π median for the *ROS* scaffold (similar in both pools). Arrows and x-axis as in C).

To investigate if the region of high divergence was associated with reduced nucleotide diversity, I looked at π across the *ROS* scaffold for each pool. As mentioned in section 3.1, recently selected alleles often result in a reduction of nucleotide diversity in and around the selected locus (due to a selective sweep or background selection). The extent of reduced π depends on the number of generations it takes for selection to remove the variability around selected loci: the quicker it is, the fewer opportunities there are for the polymorphisms around the selected locus to recombine with other haplotypes segregating in the population, and thus the larger the region of the genome with lower nucleotide diversity.

Across the whole genome, the two pools have similar distributions of π, which are highly correlated to each other (Pearson's r = 0.96). This is in accordance with the overall low *Fst* seen across the genome, indicating that pools are largely similar to each other (the nucleotide diversity within populations is not very different from that between populations, making *Fst* approach 0). In the *ROS* scaffold, the correlation between *Fst* and *π* reveals that windows with higher *Fst* have slightly lower nucleotide diversity than the rest of the scaffold (pink points in Figure 3.5); however, these values of *π* are not extreme outliers in the genome, all occurring above the 10th quantile of the genome-wide distributions (Figure 3.5). This is also represented in the profile of *π* across the *ROS* scaffold: there is a slight dip of π coincident with the *Fst* peaks, but these values are not extremely lower than the genome-wide median (Figure 3.4D).

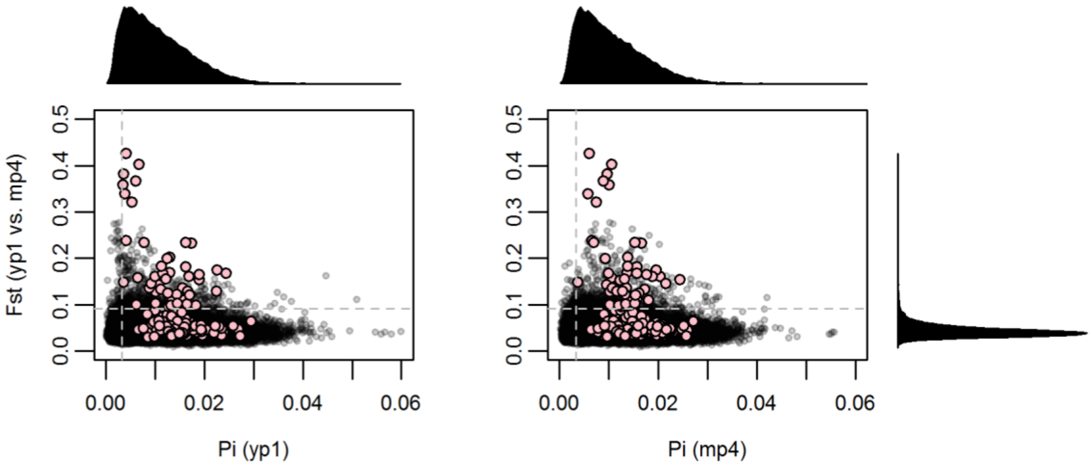


Figure 3.5 – Correlation between *Fst* and *π* for YP1 and MP4 pools across the whole genome.

The dashed vertical line is the 10th quantile of each *π* distribution (distributions shown above each plot). The horizontal dashed line is the 99th quantile of *Fst* (distribution shown on the right). The pink points are all windows in the *ROS* scaffold. Notice that windows with high *Fst* in the *ROS* scaffold have generally lower *π*, but are not extreme lower outliers of the genome-wide distributions.

To examine if the patterns of diversity observed between these two magenta and yellow pools are corroborated by comparisons made with the other sampled pools, I computed *Fst* for all pairs of those pools. Six DNA pools were collected across the flower colour cline (Table 3.1), allowing 15 pairwise combinations for which *Fst* was calculated. Because the samples are located on either side of the flower colour cline, this provides the opportunity to see how *Fst* patterns change depending on the location of each pool being compared. I divide comparisons in three broad classes: (i) between pools from the yellow side of the flower colour cline; (ii) between pools from the magenta side of the cline; and (iii) between pools from different sides of the cline. For simplification I will refer to these classes, respectively, as yellow-yellow, magenta-magenta and yellow-magenta.

In all yellow-magenta comparisons (like the one detailed for pools YP1 and MP4) the *ROS* scaffold is a clear outlier (Figure 3.6), with the maximum value of *Fst* always occurring in the top 0.2% of genome-wide *Fst* distributions. However, this is no longer the case in yellow-yellow and magenta-magenta comparisons. As before, the scaffolds containing the *PAL* and *DICH* genes are not prominent outliers in any of the comparisons.

Other outliers of high *Fst* can occur along the linkage group (and elsewhere in the genome, not shown), but many of these are not associated with any particular class of comparisons defined (e.g. grey arrow in Figure 3.6). This is likely due to a correlation between the divergence between pools and their geographic distance: the median and 99th quantile of each *Fst* distribution correlates with the distance between each pool (Figure 3.7B). For example, any *Fst* comparison including the

YP4 pool (the furthest from the hybrid zone centre) has overall higher *Fst* (highest boxplots in Figure 3.7A). This increase in the mean *Fst* for more distant pools may occur in two ways: an increase of *Fst* occurring homogeneously across all loci in the genome and/or an increase in the number of peaks of high divergence across the genome. The first explanation results in a similar dispersion of *Fst* around the mean for every comparison, whereas the second explanation results in different dispersions. Therefore, I calculated a standardized measure of dispersion for each *Fst* distribution ($coeficient\ of\ variation = \frac{standard\ deviation}{mean}$). This dispersion measure also correlates with geographic distance (Figure 3.7B), indicating that distant pools have a more heterogeneous *Fst* across the genome.

These genome-wide patterns of geographic correlation are no longer seen for the *ROS* scaffold. Several windows in this scaffold are outliers of each distribution, but only for yellow-magenta comparisons and not for yellow-yellow or magenta-magenta ones (crosses in Figure 3.7A). Consequently, the correlation of this scaffold's *Fst* with geographic distance is much weaker (Figure 3.7B). In particular, the coefficient of variation of *Fst* across this scaffold is much greater in yellow-magenta comparisons with a non-significant correlation with geographic distance (Figure 3.7B right).

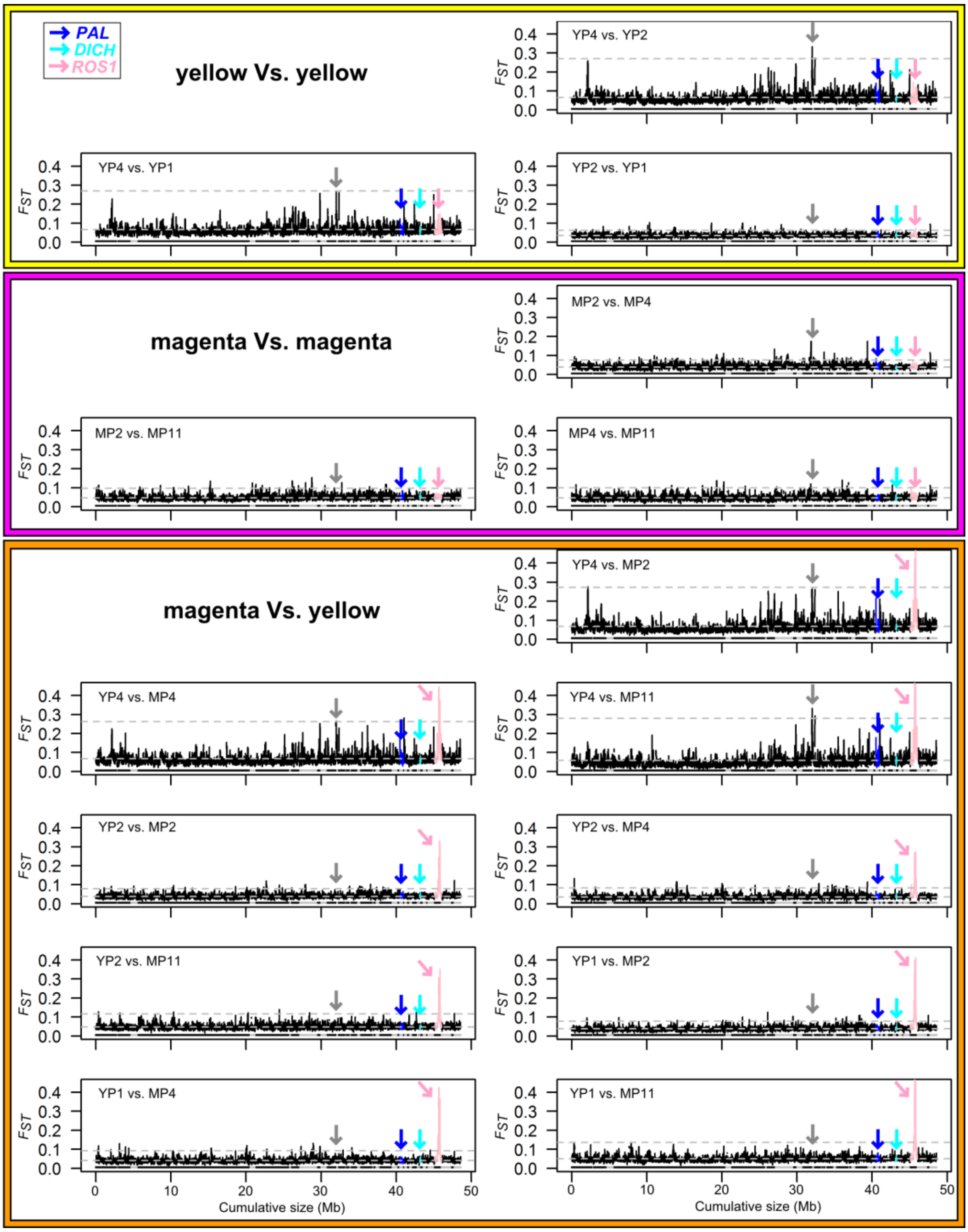


Figure 3.6 - *Fst* between pools sampled across the flower colour cline for *A. majus* linkage group 6.

All 15 pairwise comparisons are plotted for 10kb-window-averaged *Fst* (pools used in the comparison are indicated in each plot). The horizontal dashed lines represent the median and 99th quantile of each distribution. The scaffolds containing the *ROS1, PAL* and *DICH* genes are coloured as in the legend. Notice how the *Fst* signal in the *ROS1* scaffold is prominent only in yellow vs. magenta comparisons. As a contrast, the grey arrow points to a scaffold that behaves as an outlier independently of the type of comparison.

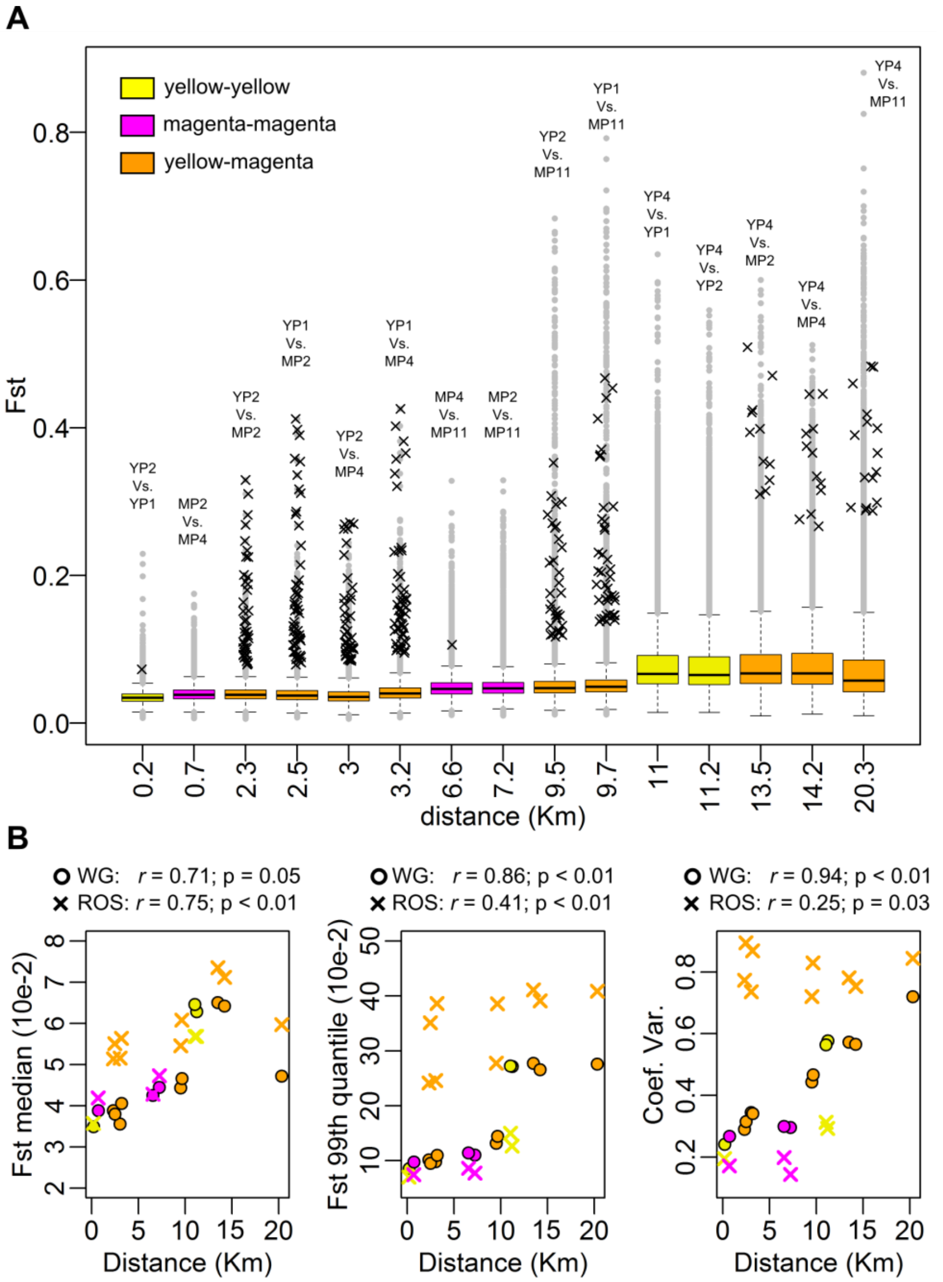


Figure 3.7 – Whole-genome distribution and patterns of geographic correlation for all pairwise Fst comparisons.

A) Boxplot of *Fst* distributions for all pairwise comparisons ordered by Euclidian geographic distance between the pools being compared. Boxes are coloured according to the type of comparison. Crosses are outlier windows in the *ROS* scaffold (*Fst* above the 99th quantile of the respective distribution). Notice that they mostly occur in yellow-magenta comparisons. Crosses are randomly shifted on the x-axis to denote their density.

B) Correlation between geographic distance and distribution measures of dispersion (left: median; middle: 99th quantile; right: coefficient of variation). The correlation (r) and p-value (p) of a Mantel test is reported above each graph for the whole genome (WG, circles) or the ROS scaffold alone (ROS, crosses). The coefficient of variation is a standardized measure of dispersion calculated as $\frac{\textit{standard deviation}}{\textit{mean}}$ of each *Fst* distribution.

3.3 Discussion

3.3.1 Patterns of divergence linked to ROS1

ROS1 is known to be genetically involved in controlling flower colour (Schwinn *et al.* 2006) and has a steep allelic cline across the *pseudomajus* x *striatum* hybrid zone (Whibley *et al.* 2006). Therefore, *ROS1* is a divergent locus between the two populations forming the hybrid zone, but the extent of divergence around this gene was unknown. Here, I characterized the divergence patterns linked *ROS1* and contrasted them with the divergence seen across the rest of the genome.

When comparing *Fst* between pools from different sides of the hybrid zone cline, the pattern of divergence across the linkage group where the *ROS* scaffold is located (LG6) is variable, but *ROS* has the highest *Fst* value (magenta-yellow comparisons in Figure 3.6). This pattern fits well with the prevalent metaphor of “genomic islands of divergence”, whereby most of the genome has low levels of divergence – the “sea-level” – punctuated by highly divergent regions – the “islands” (Figure 1.1; Turner et al. 2005; Feder and Nosil 2010; Via 2012).

The windows with the highest divergence in the *ROS* scaffold (for example, those above 99th quantile) are limited to a continuous region 180-340kb long (depending on the pairwise comparison), which corresponds to approximately 0.9-1.7 cM (the map distance is based on F2 mapping populations detailed in chapter 4, which define 1cM ≈ 200kb). The previous estimate of divergence linked to *ROS1* was based on markers 16cM and 9cM away from *ROS1*, which are non-divergent between the hybrid zone populations (Whibley *et al.* 2006). This new analysis narrows down this interval substantially. It is unreasonable to compare the size of this island with similar scans being published on other species, because this depends on the particular population history of each system under study. More useful, perhaps, would be to compare how the size of this region compares with other high *Fst* regions across the genome. Although this analysis has not been carried out systematically across the genome, there is at least one region of high *Fst* substantially bigger (Mb sized) than the peaks described here around *ROS1* (Louis Boell, pers. comm.). A systematic analysis of “divergence island” sizes would provide information on wheter *ROS1* falls within a particularly narrow “island of divergence” or if it falls within the “island” size distribution found across the genome.

A more detailed analysis of the *ROS1* region reveals that the *Fst* profile is heterogeneous around this gene (Figure 3.4B). There is one prominent peak that co-localizes with *ROS1* and two other peaks downstream of it. This could be due to noisy data. However two observations suggest that this is unlikely: (i) the result is consistent between several pairwise yellow-magenta comparisons and (ii) the peaks are not present in comparisons between pools from the same side of the cline, which instead have a flatter profile of *Fst* in the *ROS* scaffold. If the observed pattern was due to noise, yellow-magenta *Fst* comparisons should have variable peaks, but instead they consistently have the three main peaks shown in Figure 3.4B. Also, noisy data could create spurious peaks in the yellow-yellow and magenta-magenta comparisons, which is not the case (all *Fst* peaks in *ROS* are gone in such comparisons). Another source of noise could come from the fact that the reference genome is of an inbred *A. majus* stock, which can be different from the populations being studied, leading to some sequencing gaps. To avoid such biases, I only analysed windows with *P ≥ 20%*; in the particular region under focus *P = 23-89%*, which makes it unlikely that the *Fst* profile is due to an insufficient number of SNPs included in each window.

To further dissect the *Fst* pattern around *ROS1* I looked at the types of SNPs that occur in the region (Figure 3.4B). I considered that alleles occurring at a minimum frequency of 20% in both pools being compared can be considered to be “shared” between them, whereas those alleles that have a frequency difference ≥90% can be considered to be “fixed”. Classifying the SNPs in this way is helpful as it reveals their pattern of distribution along the scaffold. Several shared SNPs occur in the regions of lower *Fst* and these are interspersed by the peaks of higher divergence, which show mainly “fixed” or “private” SNPs. The presence of “shared” polymorphisms between the high *Fst* peaks suggests either gene flow that restored diversity in the region (through recombination in the hybrid zone) or shared ancestral polymorphisms (predating the divergence of *A. m. pseudomajus* and *A. m. striatum*).

Recently, it has been proposed that absolute measures of divergence that are independent from within-population diversity (π) might help distinguish between gene-flow and ancestrally-shared polymorphisms (Smith & Kronforst 2013; Cruickshank & Hahn 2014). One of these measures, d_{XY} (which measures the average number of pairwise differences between sequences from two populations) has been advocated as a way to distinguish if a region of high relative divergence (*Fst*) is due to a recent selective sweep (with no role for gene-flow) or erosion of divergence through gene-flow (Cruickshank & Hahn 2014). Unfortunately, this statistics could not be used in this work, because its calculation is not yet implemented to pooled sequence data.

A comparison with a study made in *Heliconius* butterflies suggests that the fine-scale multi-peak *Fst* linked to *ROS1* may not be uncommon (Nadeau et al. 2012). This study compared populations of *H. melpomene* with different wing pattern phenotypes. It showed that regions containing loci controlling wing colour patterns have higher *Fst* than presumably neutral regions. The pattern of *Fst* is composed of several peaks, resembling the pattern described for the *ROS1* region. These patterns may have arisen through similar evolutionary processes, particularly as these populations of *Heliconius* are also known to hybridize in nature (Baxter *et al.* 2010). A limitation when interpreting these *Fst* peaks is in knowing if each of them corresponds to an individual locus controlling the selected traits. For example, the *Fst* peaks could alternatively be due to neutral loci hitchhiking along with a single selected locus or simply due to noisy data. In the case of Nadeau et al. (2012) some of the linked *Fst* peaks relate to individual loci that are genetically characterized in both the *HmYb/Sb* and *HmD/B* regions that control wing pattern (Baxter *et al.* 2008; Ferguson *et al.* 2010; Pardo-Diaz & Jiggins 2014). This indicates that these regions remain distinct between races due to selection on wing colour polymorphisms, despite gene flow and recombination occurring in hybrid zones.

3.3.2 Nucleotide diversity in the *ROS1* region

Because *Fst* measures the divergence between populations relative to the mean diversity within each population, it might be expected that the peaks in the *ROS1* region have reduced π. However, higher *Fst* does not always imply a drop in π: for example, two populations may have a similar average number of polymorphisms, but contain a different set of haplotypes (example Figure 3.1). The nucleotide diversity measured in the hybrid zone samples was quite low: the distribution of π across the genome has a median close to zero (0.009 for YP1 and MP4) with a lower fat-tail (Figure 3.5). Finding a signal of reduced nucleotide diversity linked to *ROS1* is thus complicated since the levels of π are already quite low across the genome. Indeed, the *ROS* scaffold has no outlier *π* values correlating with high *Fst* windows (Figure 3.5). Visually, though, there seems to be a slight reduction of π in the yellow pool (YP1) co-localizing with the *ROS1 Fst* peak (pink arrow in Figure 3.4C-D). This pattern also occurs in the other yellow pools (but is less obvious in magenta pools), but I cannot confidently say that this is a significant result as none of the values are extreme low outliers of the genome-wide distributions.

The fact that the mean π is low across the genome might reflect some intrinsic features of the genome, in particular variability in recombination rates (Cutter & Payseur 2013). Regions of lower recombination (e.g. in centromeres or chromosomal inversions) often have lower diversity and these might span several megabases of sequence. The recombination rate in *ROS1* does not seem to be particularly suppressed (1 cM ~ 200 kb, detailed in Chapter 4), thus it could be that the lower π coincident with the *Fst* peaks is in fact significant. Indeed, windows with higher *Fst* in the *ROS* scaffold also have a lower π (pink points in Figure 3.5).

Other phenomena, such as an abrupt reduction of population size (e.g. bottleneck), can also result in a decrease in nucleotide diversity, but this should affect the entire genome equally, rather than just a particular locus. Therefore, assuming that the reduction of π around *ROS1* is significant, it could suggest that past selective events were involved in generating the current *Fst* peaks. For example, it could be that, in the past, positive selection for certain mutations altering flower colour, in either one or both subspecies of *A. majus*, fixed alternative haplotypes in each population (selective sweep). Alternatively, purifying selection against deleterious mutations altering the flower colour patterns might also lead to reduced diversity, because any haplotypes carrying a less fit mutation are removed from the population by selection. Intra-population diversity alone is insufficient to distinguish between selective and non-selective events, but other population genetics measures such as linkage disequilibrium and haplotype diversity can be used to clarify these scenarios (Messer & Petrov 2013). Neither of these latter statistics could be obtained from the pooled data used in this work, but future sequencing/genotyping work using individuals (rather than pools) from the hybrid zone population as well as from allopatric populations of *A. m. pseudomajus* and *A. m. striatum* might provide some clues to the evolution of *ROS* haplotypes in this species. Also, it will be important to investigate how the recombination rates across the genome in *A. majus* relate with within-population diversity and between-population divergence patterns seen across the genome, as this might help one to interpret the significance of the signals in π described here around the *ROS* locus.

3.3.3 Genomic divergence across a hybrid zone transect

The availability of six sampled pools at different distances from the *pseudomajus* x *striatum* hybrid zone allowed comparing how *Fst* across the genome correlates with the geographic distance between pools. The mean *Fst* is highly correlated with the geographic distance (Figure 3.7), i.e. pools of samples furthest away are more diverged than those closest by. Most likely this reflects population structure along the cline, implying a role for gene flow in eroding differences between pools. A similar pattern was described, for example, in *Heliconius* butterflies, whereby a significant and positive correlation was observed between *Fst* and the geographic distance between populations of the same species of *Heliconius* (Nadeau et al. 2013).

The increased divergence with geographic distance seems to be partially due to an increased number of *Fst* peaks across the genome, rather than an equal increase of *Fst* across all loci in the genome. This was evidenced by the higher *Fst* dispersion in more distant pools (Figure 3.7B), suggesting a role for gene flow across the hybrid zone in eroding some *Fst* peaks. For example, if a locus has a very broad cline across the hybrid zone (e.g. broader than the *ROS1* cline) it may be detected as an *Fst* peak when comparing the furthest pools (YP4 and MP11) but not the closest ones.

An obvious use of genomic scans of divergence is identifying loci under selection rather than those loci that are established by random drift processes. The divergence-geographic correlation described here shows the difficulty of disentangling one from the other. Some of the high divergence regions found between distant pools may reflect the incomplete mix of neutral polymorphisms between the sampled demes. This can generate a high number of false positive hits, leading to incorrectly interpret *Fst* peaks as being linked to loci under selection (Vasemägi & Primmer 2005; Via 2012). This is a general limitation of genomic scans based on outlier-based methods, since “every distribution has a tail” (Nick Barton, pers. comm.). In other words, loci falling in the tails of a distribution will always be found, since having a tail is an inherent property of distributions. An alternative to using the empirical *Fst* distribution to determine an “outlier threshold”, is to determine a threshold based on simulations that generate “neutral” distributions against which the empirical data are compared (Vasemägi & Primmer 2005). However, setting up these simulations often requires assumptions not applicable to the populations being studied and the parameters needed to run them can be difficult to estimate empirically (e.g. effective population sizes, population structure, gene flow and migration, recombination rates along the genome, demographic events). This problem is perhaps more relevant for “bottom up” approaches, that is, when phenotypes or candidate loci are unknown, which is not entirely the case in the present study (as I am focusing attention on a-priori information about *ROS1*). Still, genomic scans of divergence are useful first approaches at finding candidate loci under selection and can gain great power when complemented with other data, such as QTL or association mapping (Beaumont 2005; Via 2012).

3.3.4 Estimating nucleotide diversity and divergence from DNA pools

The measures of nucleotide diversity (π) and divergence (*Fst*) used in this chapter rely on estimating allele frequencies from populations (section 3.1). The datasets used in this work consist of whole-genome sequence data from DNA pools of several individuals sampled across the hybrid zone population. When estimating allele frequencies from pooled sequence data, the power to detect every allele in the pool is dependent on the sequencing depth *D*, leading to potential biases (Anderson *et al.* 2014). Even so, some studies suggest that sequencing a large pool of individuals is often more accurate than sequencing a few single individuals (Futschik & Schlötterer 2010; Ferretti *et al.* 2013). Usually these studies point to a minimum *D = 30-100x*, which is above the threshold being used in my analysis (*D ≥ 10x*). The reason I chose such a threshold is because at higher thresholds the number of usable windows tremendously decreases (Figure 3.3A). Re-sequencing these pools is always an option to obtain higher *D*, although it is a rather costly one. Another approach is to develop markers for a sample of SNPs, genotype the individual plants included in the pools, and compare the allele frequencies obtained with those from the pooled sequencing data. This is, in fact, underway: I have selected 50 SNPs in the *ROS* scaffold for which markers will be designed and individual plants from each pool will be genotyped, the *Fst* for each SNP calculated and compared with the estimates from pooled sequencing. Hopefully, these estimates should correlate, indicating that the pooled approach used here is reliable.

The statistics used here (*Fst* and π) were calculated as window averages, which reduces the influence of spurious errors, as discussed for the probability of obtaining wrongly fixed SNPs by chance (Willing *et al.* 2012). Indeed, the distributions of *Fst* using *D ≥ 10x* or *D ≥ 20x* were not very different, particularly for window *P ≥ 20%* (Figure 3.3B). A strong point of this dataset is that it comprises several pools, which have been compared in multiple pairwise comparisons that agreed in the patterns observed around *ROS1*. This suggests that sequencing of DNA pools can be a valuable and relatively affordable approach (compared to sequencing of individuals) to obtain diversity measures across the genome in wild populations.

In summary, this chapter demonstrates that the region linked to *ROS1* is highly divergent between samples from opposite flanks of the hybrid zone, compared to the remainder of the genome. At a fine-scale, the profile of divergence is heterogeneous around *ROS1*, with at least 3 individual *Fst* peaks characterised by a higher number of “fixed” polymorphisms than in the surrounding regions, where they are mostly absent. A question arising from this analysis is whether the individual peaks correspond to individual loci controlling the flower colour differences seen between the two *Antirrhinum* subspecies. This question is addressed in the next chapter by genetic and molecular mapping of loci linked to *ROS1*.

4 Genetic mapping of loci linked to divergence peaks

The peaks of high *Fst* linked to *ROS1* suggest that other loci may be divergent between *A. m. pseudomajus* and *A. m. striatum*. Previous genetic experiments suggest that a locus linked to *ROS1*, named *ELUTA* (*EL*), controls other aspects of flower colour, making it a likely candidate. However, *EL* has never been fine-mapped, thus its location relative to *ROS1* is unknown. Using several genetic strategies to generate recombinants, I show that *ROS* and *EL* are two separate loci ~0.5 cM apart. Using these recombinants, I narrowed the location of *EL* to a ~50 kb interval, located downstream of the *ROS1* gene. This interval precicely coincides with one of the divergence (*Fst*) peaks downstream of the peak containing *ROS1* (described in the previous chapter), supporting that this locus is under selection in the *pseudomajus* x *striatum* hybrid zone. Finally, I show that within the *ROS* locus itself, there are different regions linked to *ROS1* that control the intensity of magenta pigment in the flowers. Importantly, I show how a detailed genetic analysis improves the interpretation of divergence patterns across genomes, which in this case precisely pinpoint the loci involved in a reproductive barrier between hybridizing populations.

4.1 Introduction: genetics of *ROS* and *EL*

The genetic and molecular characterization of *ROS* was based on several lines available in the *A. majus* collection at the John Innes Centre (Figure 4.1). These include the canonical “wild type” *JI7*, with a full magenta pigment and two lines with reduced floral pigmentation resulting from mutations in the *ROS* locus, named *rosea*^{dorsea} and *rosea*^{colorata} (Figure 4.1A-C). In the *rosea*^{dorsea} line, the magenta pigment is restricted to the outer epidermis of the dorsal part of the flower lobes. In *rosea*^{colorata}, the pigment occurs in a ring at the base of the corolla tube and in the inner epidermis of the lobes.

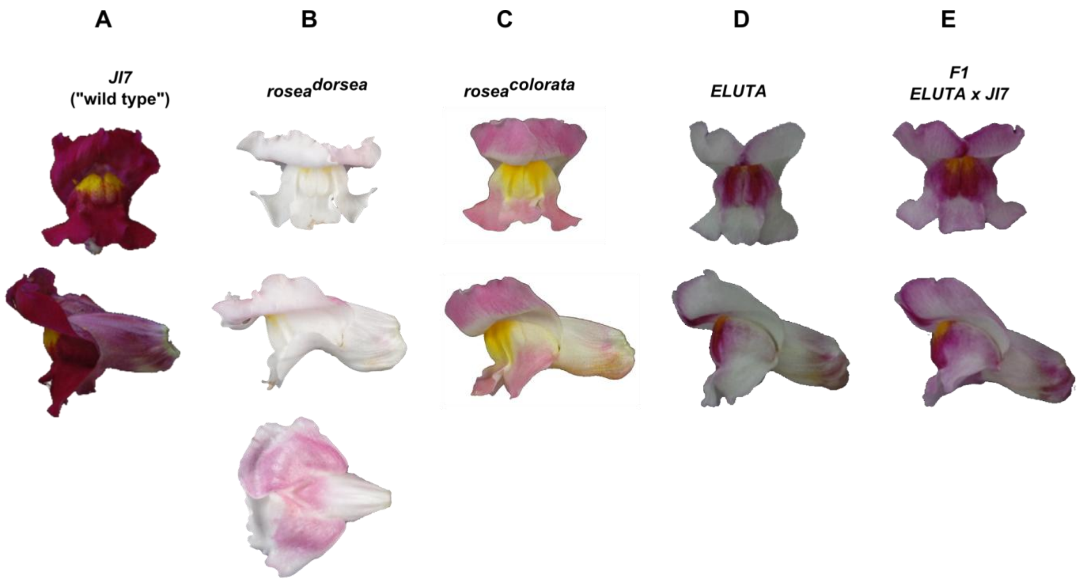


Figure 4.1 – Phenotypes of *Antirrhinum majus* lines available at the John Innes Centre. The front (top), side (middle) and dorsal (bottom) views of the flowers is shown. The *JI7* line (A) is the canonical "wild type" with magenta flowers. *rosea*^{dorsea} (B) and *rosea*^{colorata} (C) each have a different mutant allele in the *ROS* locus; both mutations are recessive to the *JI7 ROS*. The *ELUTA* line (D) carries a semi-dominant allele of the *EL* locus, which restricts the pigmentation in the flowers; the heterozygote *ELUTA x JI7* (E) has a less severe phenotype than the homozygous line.

The *ROS* locus includes two tandemly duplicated *MYB*-like genes: *ROS1* (focused on in the previous chapter) and *ROS2* (Schwinn *et al.* 2006). These genes were identified based on their sequence similarity to other *MYB*-like transcription factors that regulate anthocyaning production (namely *C1* in maize and *AN2* in petunia; Paz-Ares et al. 1987; Quattrocchio et al. 1999) and the fact that they were differentially expressed between the different *rosea* lines of *A. majus* (Figure 4.1). In the magenta *JI7* line, only *ROS1* is expressed, suggesting it is sufficient for the production of pigment. In *rosea*^{dorsea}, *ROS1* has very low expression and this correlates with a reduction of pigment. Finally, *rosea*^{colorata} has a loss-of-function allele of *ROS1* but it expresses *ROS2*. Taken together, the analysis of these different *A. majus* lines suggests that *ROS1* has a major role in the production of the magenta pigment of flowers (being sufficient for the floral phenotype of *JI7*), but *ROS2* can also have a minor contribution to this trait (as seen in *rosea*^{colorata}) (Schwinn *et al.* 2006).

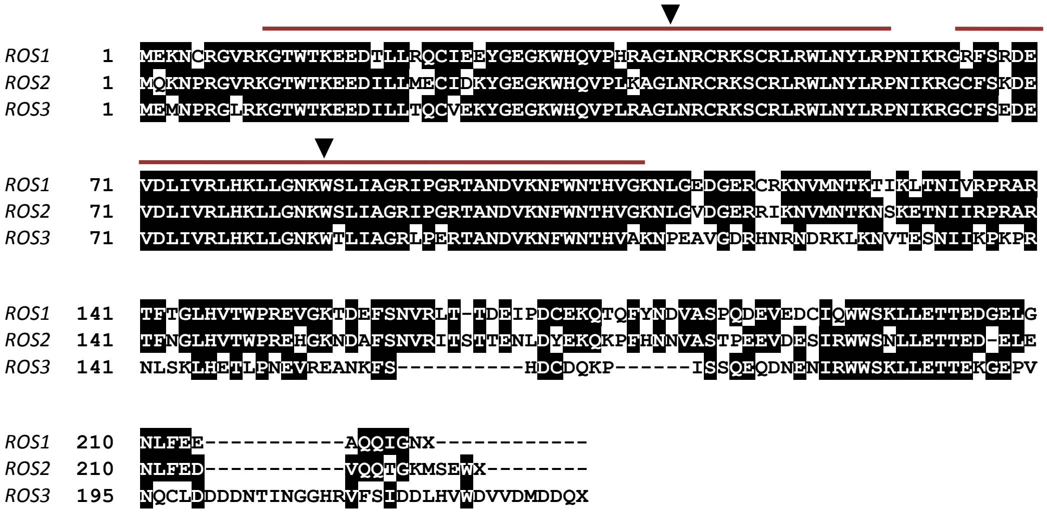

ROS1 1 MEKNCRGVRKGTWTKEEDTLLRQCIEEYGEGKWHQVPHRAGLNRCRKSCRLRWLNYLRPNIKRGRFSRDE
ROS2 1 MQKNPRGVRKGTWTKEEDILLMECIDKYGEGKWHQVPLKAGLNRCRKSCRLRWLNYLRPNIKRGCFSKDE
ROS3 1 MEMNPRGLRKGTWTKEEDILLTQCVEKYGEGKWHQVPLRAGLNRCRKSCRLRWLNYLRPNIKRGCFSEDE
ROS1 71 VDLIVRLHKLLGNKWSLIAGRIPGRTANDVKNFWNTHVGKNLGEDGERCRKNVMNTKTIKLTNIVRPRAR
ROS2 71 VDLIVRLHKLLGNKWSLIAGRIPGRTANDVKNFWNTHVGKNLGVDGERRIKNVMNTKNSKETNIIRPRAR
ROS3 71 VDLIVRLHKLLGNKWTLIAGRLPERTANDVKNFWNTHVAKNPEAVGDRHNRNDRKLKNVTESNIIKPKPR
ROS1 141 TFTGLHVTWPREVGKTDEFSNVRLT-TDEIPDCEKQTQFYNDVASPQDEVEDCIQWWSKLLETTEDGELG
ROS2 141 TFNGLHVTWPREHGKNDAFSNVRITSTTENLDYEKQKPFHNNVASTPEEVDESIRWWSNLLETTED-ELE
ROS3 141 NLSKLHETLPNEVREANKFS----------HDCDQKP------ISSQEQDNENIRWWSKLLETTEKGEPV
ROS1 210 NLFEE-----------AQQIGNX------------
ROS2 210 NLFED-----------VQQTGKMSEWX--------
ROS3 195 NQCLDDDDNTINGGHRVFSIDDLHVWDVVDMDDQX

Figure 4.2 – Protein alignment of *ROS1, ROS2* and *ROS3*.

The protein sequences were translated from the predicted CDS of each gene, based on *A. majus JI7* genome sequence. Lines above the alignment denote the two conserved *MYB* DNA-binding domains. Triangles indicate the conserved position of the two introns in the genomic sequence.

ROS1 and *ROS2* are tightly linked: the first exon of *ROS2* is just 5.5kb downstream of the last exon of *ROS1*. The two genes are highly similar (77% similarity at the protein level) with most differences in the C-terminus; in the *MYB* domain of the protein there are only 9/99 amino-acid differences (91% similarity) (Figure 4.2). By analysing the *ROS* scaffold, I found a third *ROS*-like gene, with high similarity to both *ROS1* and *ROS2*. This gene is located 6.6kb downstream of *ROS2* and was named *ROS3*. Its function in regulating flower colour is, at this point, unknown. For nomenclature purposes, I will use a number suffix to refer to each of these genes individually (*ROS1, ROS2, ROS3*), whereas the entire locus will be referred to as *ROS*.

The described lines of *A. majus* allow the functionality of *ROS* in different phenotypes to be looked at. The *ROS* allele conferring full magenta in *JI7* is dominant over the recessive *ros*^{dor} and *ros*^{col} alleles (Schwinn *et al.* 2006). Therefore, the *rosea*^{dorsea} mutant can be used in complementation tests with *A. m. pseudomajus* and *A. m. striatum*, to find if they carry a functional allele of *ROS* (Figure 4.3; Whibley et al. 2006). The F1 progeny between *A. m. pseudomajus* and *ros*^{dor} has full magenta flowers (Figure 4.3-iv), suggesting that this subspecies carries a dominant functional allele of *ROS*, similar to *JI7*. Conversely, *A. m. striatum* is unable to complement the *rosea*^{dorsea} phenotype (Figure 4.3-i) suggesting that it carries a recessive loss of function allele in *ROS*.

A cross between the full magenta line, *JI7*, and the subspecies, reveals the effect of an otherwise cryptic genetic locus. While the F1 progeny of *A. m. pseudomajus* x *JI7* results in full magenta progeny (identical to both parents; Figure 4.3-v) that of *A. m. striatum* x *JI7* results in a restricted pigmentation pattern (Figure 4.3-ii). This effect is due to *ELUTA* (*EL*), a semi-dominant locus that alters the pigmentation pattern in the flowers (Schwinn *et al.* 2006; Whibley *et al.* 2006). This phenotype (which I will refer to as “Eluta”) is characterized by an overall reduction of magenta pigment, which becomes largely restricted to the base of the flower tube and the portion of the ventral lobe where the ventral and lateral petals fuse (Figure 4.1D). This locus was initially characterized based on a line of *A. majus* with this phenotype (Figure 4.1E), but later shown to also have functionally distinct alleles in different *Antirrhinum* species (Schwinn *et al.* 2006; Whibley *et al.* 2006). The semi-dominance effect of this locus is visible in the *ELUTA* line of *A. majus*: a heterozygous *ELUTA* x *JI7* individual has a less restricted pigmentation than the homozygous *ELUTA* individual (compare Figure 4.1 D and E).

Analysis of phenotypic segregation in F2s reveals that *ROS* and *EL* segregate together. For example, the full magenta phenotype is rarely recovered in F2s from an *ELUTA* x *rosea*^{dorsea} cross; instead a 1:3 proportion of white:Eluta phenotypes is observed, indicating genetic linkage between *ROS* and *EL* (Baur 1911, 1912; Schwinn *et al.* 2006). *ROS* and *EL* also show genetic epistasis: the effect of *EL* depends on which *ROS* allele is present in the background. For example, *A. m. striatum* carries a semi-dominant allele of *EL* (Figure 4.3-ii), but the Eluta phenotype is not visible in the *A. m. striatum* parent because it carries a recessive allele of *ROS* (Figure 4.3-i). Therefore, the genotype in each locus cannot be determined based on an individual’s phenotype alone, but requires the analysis of crosses with *A. majus* lines of known genotype (Figure 4.3).

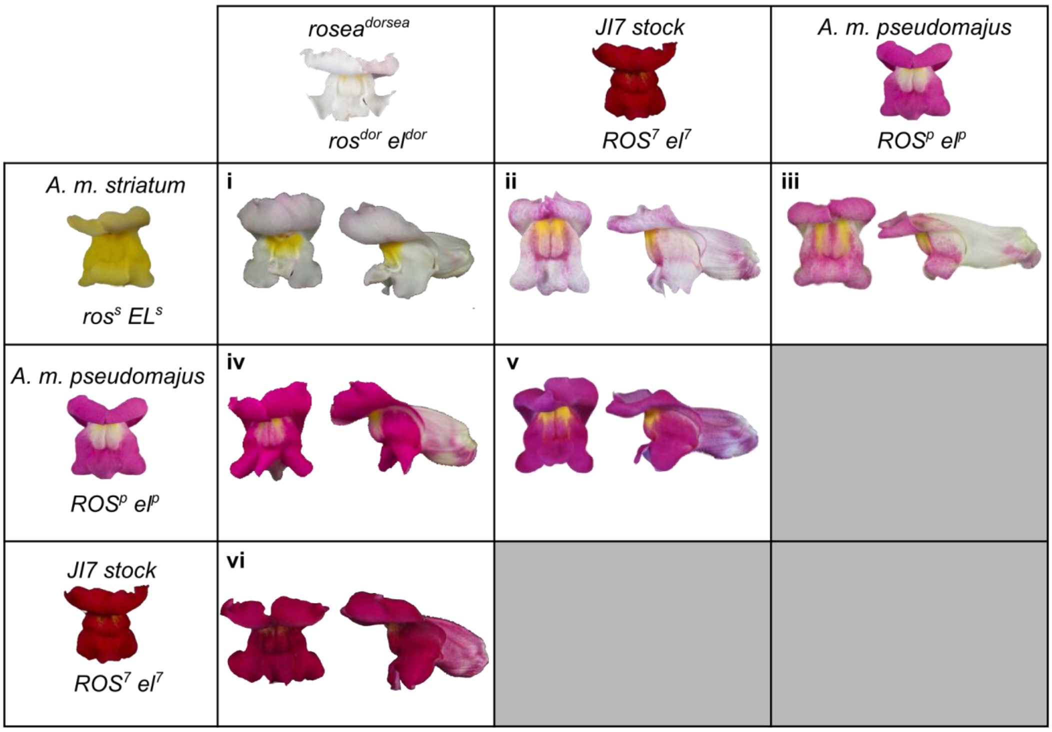


Figure 4.3 – F1 phenotypes of complementation tests involving *A. m. pseudomajus* and *A. m. pseudomajus*.

The matrix shows the F1 phenotype of crosses between each subspecies and either *rosea*^{dorsea} or *JI7*. The cross to *rosea*^{dorsea} and to *JI7* is informative with regards to the genotype in *ROS* and *EL*, respectively. If a plant carries a dominant allele of *ROS*, then the F1 x *rosea*^{dorsea} should be pigmented. If a plant carries a dominant allele of *EL*, then the F1 x *JI7* should have a restricted pigment. Deduced genotypes of each plant are indicated, with uppercase indicating dominant alleles and lowercase recessive ones. The superscript in each genotype refers to the allele carried by each individual (p – *pseudomajus*; s – *striatum*; 7 – *JI7*; d - *rosea*^{dorsea}). Adapted from Whibley (2006).

An *F1* cross between *A. m. pseudomajus* and *A. m. striatum* reveals the phenotypic consequence of the *ROS*-*EL* genetic interaction in the context of the hybrid zone between these subspecies. While the parental species have magenta and non-magenta phenotypes respectively, the F1 individuals have an Eluta phenotype (Figure 4.3-iii), which is widely observed in the *pseudomajus* x *striatum* hybrid zone population. This is due to F1 hybrids between the subspecies being heterozygous for *ROS* as well as *EL*. The segregation of phenotypes in an F2 progeny from this cross confirms the linkage of the two loci in these subspecies (section 4.2; Whibley 2004).

EL has yet to be cloned, and it has so far proven difficult to map due to its tight linkage to *ROS*. Schwinn et al. (2006) did not find recombinants between *ROS* and *EL*, whereas Whibley et al. (2006) found 2 recombinants in a test-cross progeny of 1300 plants (detailed below). The latter work suggests that *ROS* and *EL* are separate loci (rather than alleles of the same locus), but also demonstrates that obtaining recombinants requires analysis of large numbers of progeny in mapping populations.

The extent of yellow colour in the lobes of the flowers, which is conferred by aurone pigments, is largely controlled by a locus named *SULF* (Whibley 2004; Whibley *et al.* 2006). This locus has yet to be cloned, but it is unlinked to *ROS-EL*. Because the yellow colour is not the focus of this work, I will not consider its contribution to the floral phenotype in this study. However it should be noted that this is an important feature that, together with the magenta pigment, distinguishes between the two subspecies of *A. majus* studied here.

In this chapter, I will refer to the phenotypes of the flowers as magenta (like *JI7* and *A. m. pseudomajus*), Eluta (like F1 *pseudomajus* x *striatum* and *JI7* x *striatum*) and non-magenta (like *rosea*^{dorsea} and *A. m. striatum*). For the genetic analysis we need to determine haplotypes; that is, the combination of alleles in the two adjacent *ROS* and *EL* loci. When referring to such haplotypes I use the notation <u>*ROS EL*</u> (the underline denotes the physical linkage between them). Dominant (or semi-dominant) and recessive alleles are written in uppercase and lowercase letters respectively, with a superscript referring to the origin of the allele. Examples relevant throughout the work are: $\underline{ROS^{7}\ el^{7}}$ for a *JI7* haplotype; $\underline{ROS^{p}\ el^{p}}$ for a *pseudomajus* haplotype, $\underline{ros^{s}\ EL^{s}}$ for a *striatum* haplotype; and $\underline{ros^{dor}\ el^{dor}}$ for a *rosea*^{dorsea} haplotype. To simplify the notation, I sometimes omit the superscript, but the uppercase/lowercase notation always indicates which kind of allele is being referred to. Because the *JI7* and *pseudomajus* haplotypes are genetically equivalent to each other (both are $\underline{ROS\ el}$), I may use the superscript in the form of $\underline{ROS^{7/p}\ el^{7/p}}$, meaning the haplotype originates from either one of those lines. Finally, to indicate the diploid genotype of an individual I will separate the haplotypes by a "/" character. For example, a heterozygous F1 *pseudomajus* x *striatum* is $\underline{ROS^{p}\ el^{p}}$ / $\underline{ros^{s}\ EL^{s}}$.

4.2 Results

4.2.1 Mapping *EL* from an F2 of *A. m. pseudomajus* x *A. m. striatum*

As a first approach to find recombinants between *ROS* and *EL,* I looked for discrepancies between the *ROS1* genotype and the colour phenotype in F2 progenies of *A. m. pseudomajus* x *A. m. striatum*. If no recombination between *ROS* and *EL* occurs, then the genotype of *ROS1* should predict the phenotype as follows: homozygotes for the *pseudomajus* allele ($\underline{ROS\ el}$ / $\underline{ROS\ el}$) should be magenta; homozygotes for the *striatum* allele ($\underline{ros\ EL}$ / $\underline{ros\ EL}$) should be non-magenta; heterozygotes ($\underline{ROS\ el}$ / $\underline{ros\ EL}$) should be Eluta.

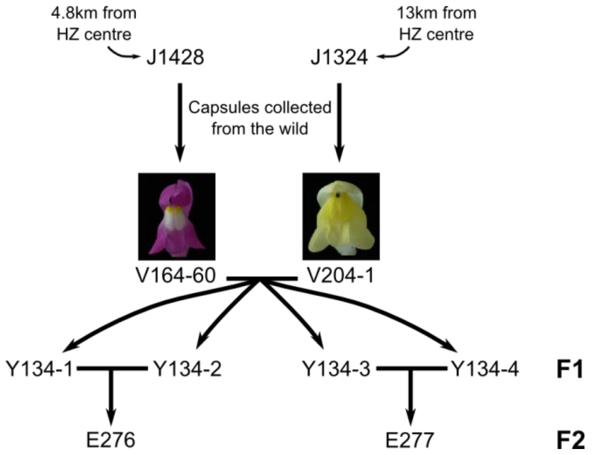


Figure 4.4 – Crosses to generate *A. m. pseudomajus* x *A. m. striatum* F2 families. Capsules were collected from wild individuals distant from the hybrid zone centre: individual J1428 was 4.8km on the *pseudomajus* side; individual J1324 was 13km on the *striatum* side. These were sown and two individuals with typical phenotypes of each subspecies (shown in photos) were selected to generate an F1 (family Y134). Four individuals from this F1 were inter-crossed to produce segregating F2s (families E276 and E277).

To generate segregating F2 families between the two subspecies, individuals originating from the wild were used. Capsules from wild *A. m. pseudomajus* and *A. m. striatum* individuals were collected and the seeds grown in the glasshouse. Individuals with typical *pseudomajus* and *striatum* phenotypes from each capsule were intercrossed to produce F1 progeny (Figure 4.4). Because wild *A. majus* are self-incompatible, the F2 generation was produced by crossing F1 siblings. Two families were produced, of 96 individuals each. From these 192 plants, 161 were successfully genotyped with two markers in the *ROS1* gene which, together, distinguish between the two subspecies' alleles (markers 5 and 6 in Table 2.1).

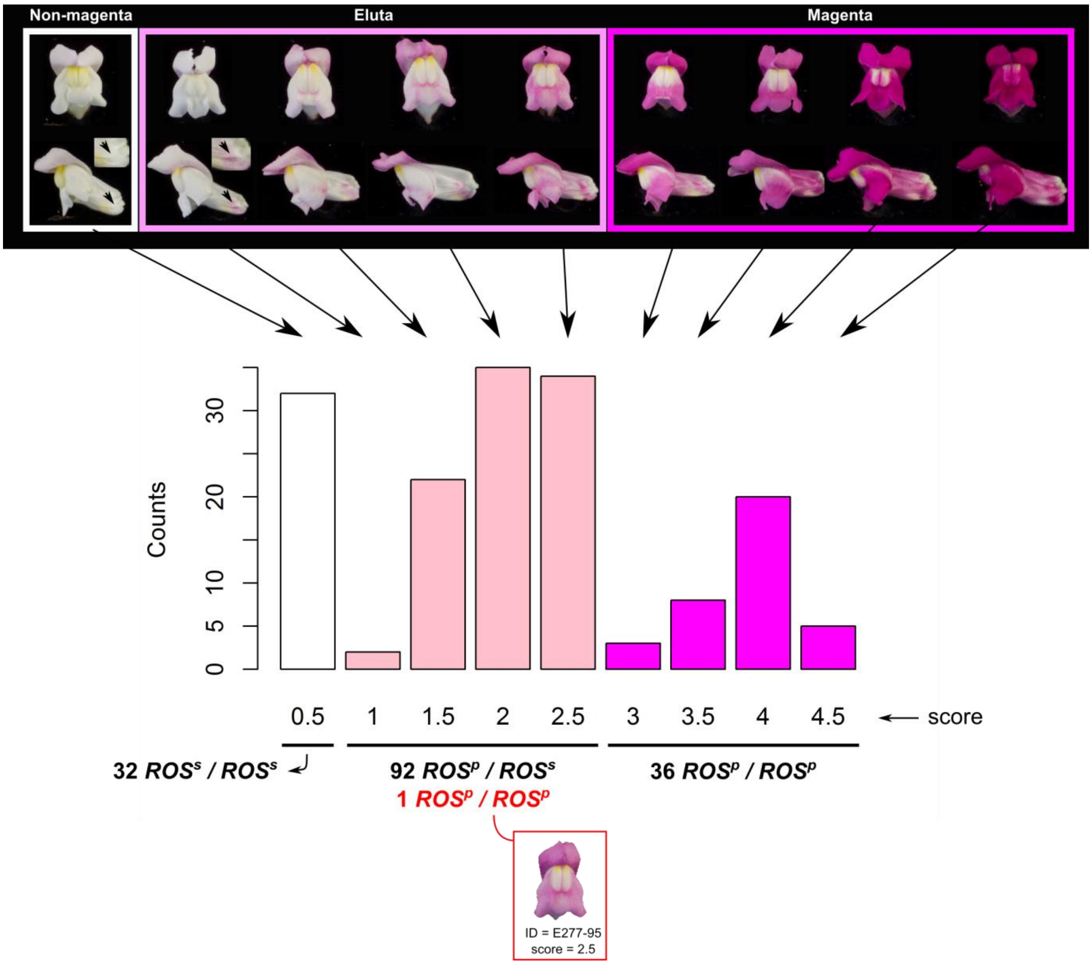


Figure 4.5 – Results from *A. m. pseudomajus* x *A. m. striatum* F2.
161 individuals from two families (E276 and E277 in Figure 4.4) were scored for their magenta phenotype. The score, adapted from Whibley (2004), varies between 0.5-5, with increasing values corresponding to increasing pigmentation. The barplot shows the counts of each phenotype score, with representative photographs for each shown above. Notice that individuals with score 1 are distinguishable from those with score 0.5 by a subtle presence of pigment in the base of the tube (arrow and inset). The individuals were genotyped for *ROS1*¸ with markers that distinguish the two subspecies alleles. The numbers of each genotype are shown below the barplot for the three phenotypic classes. The superscript refers to the origin of the allele: p – *pseudomajus*; s – *striatum*). One exceptional genotype (in red font and boxed) was observed with an Eluta phenotype.

The magenta phenotype of these F2 plants was visually scored using a system similar to Whibley (2004). The score varies between 0.5 and 4.5, with increasing values corresponding to increased pigmentation in the flowers (Figure 4.5). These scores can be divided into the phenotype classes defined in the previous section:

non-magenta (0.5 ≤ score < 1), Eluta (1 ≤ score < 3) and magenta (3 ≤ score ≤ 4.5). The respective segregation of these phenotypes in the F2 progeny was 32 : 93 : 36, which is not significantly different from the 1 : 2 : 1 ratio expected for a single Mendelian locus (Chi-square test: $\chi^2 = 4.0807$; d.f. = 2; $p = 0.13$). However, this ratio is also not significantly different (and in fact is a better fit) to the segregation expected for two unlinked loci with an epistatic interaction between them, which is 3 : 9 : 4 (Griffiths *et al.* 1993) (magenta : Eluta : non-magenta; Chi-square test: $\chi^2 = 0.6232$; d.f. = 2; $p = 0.73$).

To distinguish between the two hypotheses - that *ROS* and *EL* are linked or unlinked - the F2 progeny was genotyped with molecular markers in the *ROS1* gene. If the two loci are unlinked, the *ROS1* genotype should not always predict the phenotype. For example, to have magenta flowers an individual could either be homozygous or heterozygous for $ROS1^p$, in a proportion of 1:2 respectively (as long as it is recessive el^s/el^s). Contrary to this, and compatible with tight linkage, the genotyping results revealed an almost perfect association between *ROS1* genotype and phenotype (Figure 4.5): individuals homozygous for the *pseudomajus* allele ($ROS1^p$ / $ROS1^p$) had magenta flowers; individuals homozygous for the *striatum* allele ($ros1^s$ / $ros1^s$) had non-magenta flowers; and heterozygous individuals ($ROS1^p$ / $ros1^s$) had Eluta flowers. There was only one exception: individual E277-95 was homozygous for the *pseudomajus* $ROS1^p$ allele, but had an Eluta phenotype characteristic of a heterozygote. This was interpreted to be the consequence of recombination between *ROS* and *EL*, resulting in a $\underline{ROS^p\ EL^s}$ haplotype heterozygous with a $\underline{ROS^p\ el^p}$ haplotype. This being true, there was 1 recombination event in 322 meiotic products analysed (two from each of the 161 plants genotyped), which gives an estimate of $\frac{1}{322} \times 100 = 0.3cM$ separating the two loci. This confirmed the tight linkage between *ROS* and *EL*, but for fine-mapping the two loci a larger collection of recombinants was required, as well as molecular markers linked to *ROS1* to map the recombination points.

The Eluta phenotype in these families is quite variable in comparison to the *A. majus ELUTA* line (Figure 4.1D). For example, some individuals with Eluta phenotype have almost no visible pigment in the flowers (score 1), only being distinguished from non-magenta individuals by a coloration at the base of the flower tube (Figure 4.5). This is possibly due to the segregation of other modifiers of pigment in the background of these subspecies (further discussed in chapter 6).

4.2.2 Fine-mapping recombination points between *ROS* and *EL*

To produce larger numbers of recombinants between *ROS* and *EL* two main approaches were used, one based on genotype and one based on phenotype (Figure 4.6). In the first approach, I genotyped an F2 population segregating for a *striatum* haplotype ($\underline{ros^{s}\ EL^{s}}$) and a *JI7* haplotype ($\underline{ROS^{7}\ el^{7}}$), with markers linked to *ROS1*. These markers distinguish between the *JI7* and *striatum* alleles. I then identified recombination points and associated them with changes in the phenotype (genotypic screening in Figure 4.6). The other approach consisted of screening test-cross progenies, which allowed identification of recombinants by phenotype, and these were later genotyped to identify recombination points (phenotypic screen in Figure 4.6).

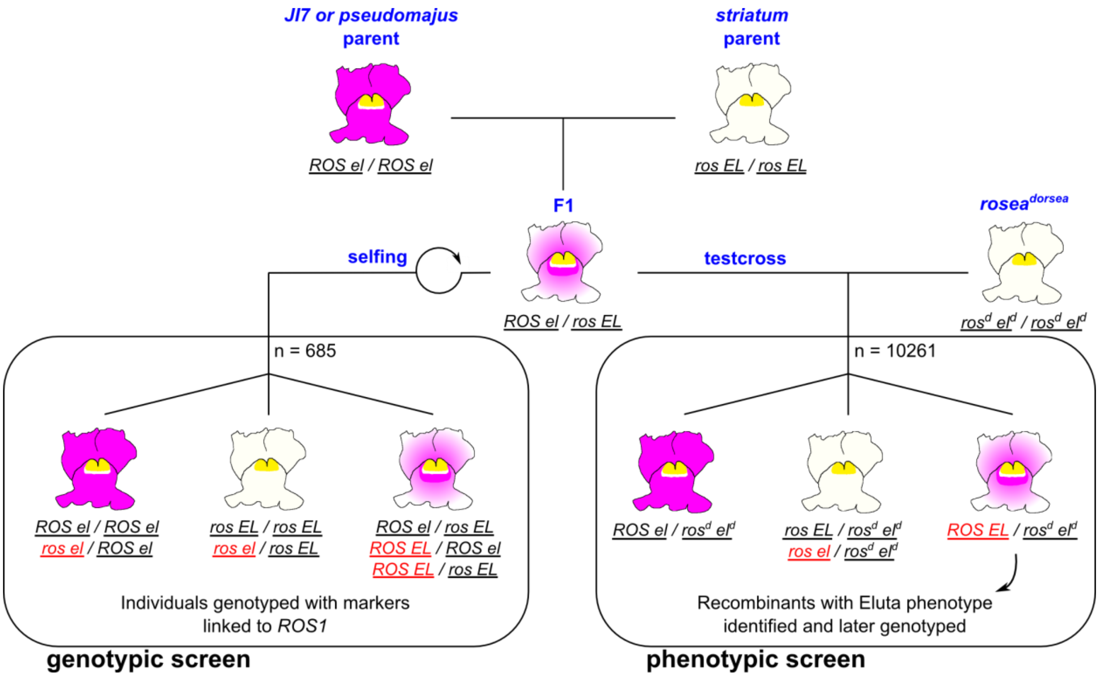


Figure 4.6 – Genetic strategies to screen recombinants.

Heterozygous ROS el/ros EL individuals carrying the *A. m. striatum* and either the *JI7* or *A. m. pseudomajus* haplotypes were used to produce recombinants between the two loci. One strategy consisted of sowing an F2 that was genotyped with several markers linked to *ROS1* (genotypic screen, on the left). The other strategy consisted of sowing large numbers of test-cross progenies which were screened based on phenotype (phenotypic screen, on the right). The test-cross progeny consisted of crossing the heterozygote with the *rosea*^{dorsea} mutant, which has a double recessive haplotype (ros el), therefore not contributing to the pigmentation of the flowers. Magenta and non-magenta phenotypes are expected due to segregation of the parental haplotypes, whereas an Eluta phenotype is expected if a recombination occurs between *ROS* and *EL.* The expected genotypes in each case are indicated below the flower pictograms with uppercase letters denoting dominant alleles and lowercase recessive ones (the two haplotypes are separated by "/"). Recombinant haplotypes are in red. In the genotypic screen, some individuals with recombinant haplotypes have a phenotype indistinguishable from that of a non-recombinant.

4.2.2.1 Mapping recombination points associated with ELUTA

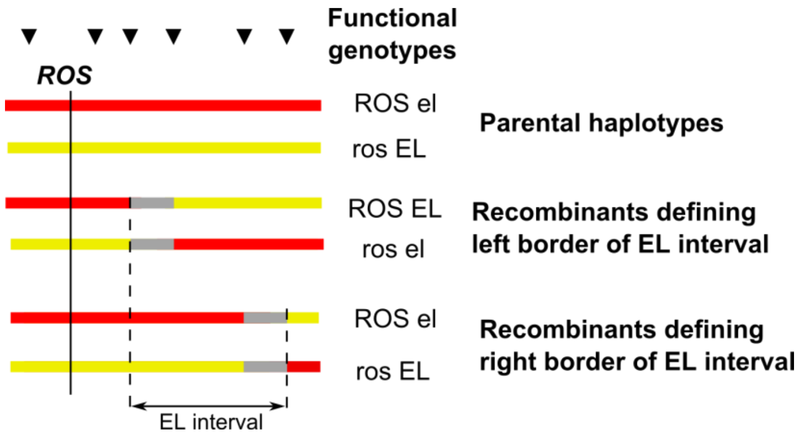
To fine-map *EL*, I genotyped 685 *JI7 x striatum* F2 individuals (genotypic screen Figure 4.6) with 4 markers that span a distance of ~398kb in the *ROS* scaffold. This interval extends ~75kb upstream and ~323kb downstream of *ROS1*. A total of 25 recombinant haplotypes were detected and these were subsequently genotyped with another set of 13 markers to narrow down the recombination points. For reference, the results from this mapping are included in a full panel of recombinants (Figure 4.15A), but here I will focus only on key recombinants that allowed mapping *EL*.

			ROS1 ↑								
Marker Number	**3**		**5**		**16**		**29**		**32**		**38**
Distance, kb		74.397		96.028		62.620		15.119		149.494	
Nr recombinations		5		6		4		0		10	
Map distance, cM (1370 meiotic products)		0.36		0.44		0.29		NA		0.73	
Kb/cM		**207**		**218**		**216**		NA		**205**	

Figure 4.7 – Map distances between markers around *ROS1*.
An F2 family segregating for *striatum* and *JI7 ROS-EL* haplotypes was genotyped with several markers around *ROS1*. The map distance (cM) between each marker was calculated as the percentage of recombination events in the total number of meiotic products (2 haplotypes obtained from each of 685 individuals). The number of base-pairs corresponding to one map distance was calculated by using the distance between markers in the *A. majus* reference genome. No recombination was observed between markers 29 and 32, and therefore calculating a map distance was not applicable (NA). Marker numbers correspond to Table 2.1.

Before mapping *EL* I checked to see if recombination is homogeneous across the region linked to *ROS1*. Therefore, I determined the physical distance corresponding to one map distance (1 cM) for this region. I calculated the percentage of recombination between 6 consecutive markers and divided it by their distance in the *A. majus* genome (Figure 4.7). On average 1cM ≈ 211kb and the recombination rate is fairly constant across the region, with estimates varying between 205-218 kb/cM. This suggests that recombination is not suppressed around *ROS1* and therefore I use an approximation of 1cM = 200kb.

To identify which of the recombinant haplotypes found by the genotypic screen resulted in a recombination point between *ROS* and *EL* (as opposed to recombination points that do not uncouple the parental *ROS* and *EL* alleles), I looked for discordances between the *ROS1* genotype and the expected phenotype of the recombinants. For example, if an individual is homozygous $ROS1^{7}$ / $ROS1^{7}$ but has an Eluta phenotype then it must carry a dominant allele EL^{s} (i.e. a $\underline{ROS^{7}\ EL^{s}}/\underline{ROS^{7}\ el^{7}}$ instead of $\underline{ROS^{7}\ el^{7}}/\underline{ROS^{7}\ el^{7}}$). However, some recombinants may have a phenotype similar to non-recombinants (Figure 4.6). For example, a homozygote with a *striatum* haplotype ($\underline{ros^{s}\ EL^{s}}$ / $\underline{ros^{s}\ EL^{s}}$) will be non-magenta, but so will a heterozygote with a *striatum* haplotype and a $\underline{ros^{s}\ el^{7}}$ recombinant haplotype ($\underline{ros^{s}\ el^{7}}$ / $\underline{ros^{s}\ EL^{s}}$). This is because both of them lack the ROS^{7} allele that allows producing pigment and therefore *EL* does not contribute to the pigmentation pattern of these flowers. Because of this, some recombinants were either selfed to produce an F2 or crossed to $rosea^{dorsea}$ or *JI7* stocks (see complementation tests in Figure 4.3). One recombinant never produced flowers, so it was excluded from this analysis. I thus determined the *ROS* and *EL* genotypes for 24 out of 25 recombinants. The interval containing *EL* could be determined by relating the recombination points that result in the uncoupling of the parental *ROS* and *EL* alleles (i.e. resulting in *ros el* or *ROS EL* haplotypes) and those recombination points that do not (explained in Figure 4.8).



Functional genotypes
ROS
ROS el
ros EL
Parental haplotypes
ROS EL
ros el
Recombinants defining left border of EL interval
ROS el
ros EL
Recombinants defining right border of EL interval
EL interval

Figure 4.8 – Schematic view of the strategy for determining the interval containing *EL*. The horizontal coloured lines represent a portion of a chromosome, with colours denoting the origin of the genotype (*striatum* in yellow; *JI7* in red), which is determined by diagnostic molecular markers (triangles). The grey interval represents a portion of the chromosome of undetermined origin (flanked by two markers with different genotypes). The recombinants can be used to determine an interval containing *EL*. First, the genotype in *ROS* and *EL* of each recombinant is determined genetically (e.g. from complementation tests). Recombination points that uncouple the parental *ROS* and *EL* alleles define the left border of an interval containing *EL*. Recombination points that do not change the parental *ROS* and *EL* combination define the right border of that interval.

There were 9 recombination events resulting in uncoupled parental *ROS* and *EL* alleles, mapping them $\frac{9}{1370} \times 100 = 0.66\, cM$ apart. This figure is higher than the previously estimated 0.3 cM, which is likely due to the different sample sizes of the two screens (322 vs. 1370 meiotic products). Uncoupling of the two loci always occurred with recombination points mapping downstream of *ROS1* (Figure 4.9A). The most distant of these recombination points (thus closest to *EL*) was 111.150 kb away from *ROS1* (marker 21 in Figure 4.9A). This marker thus defined the left border of an interval containing *EL*. Recombination points 191.498 kb downstream of *ROS1* no longer uncoupled *ROS* and *EL* (marker 35 in Figure 4.9A), thus defining a right border for the *EL* interval. This analysis of recombinants allowed mapping *EL* to an 80 kb interval downstream of the *ROS1* gene.

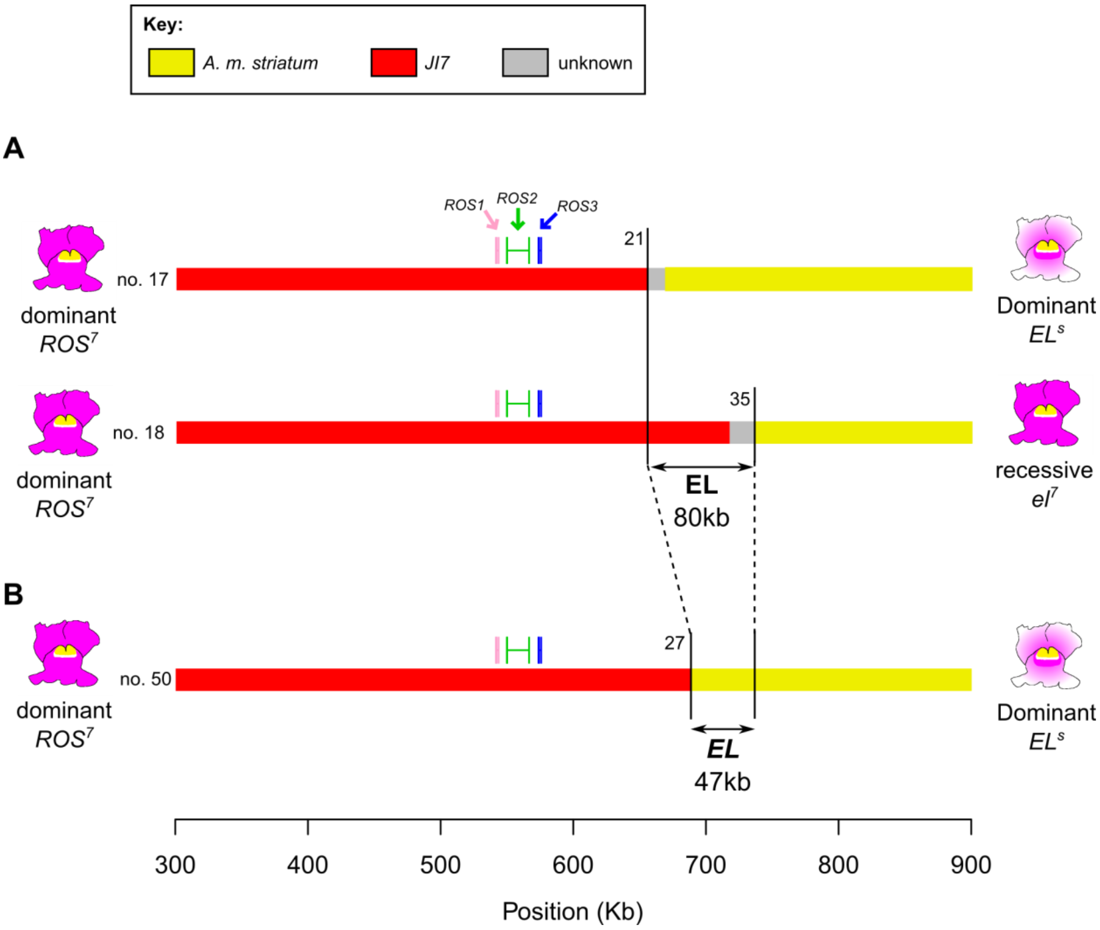


Figure 4.9 – Mapping the *EL* locus.

Recombinant haplotypes around the *ROS* locus were genotyped with molecular markers in the *ROS* scaffold. Haplotypes are represented by horizontal lines; colours are: red - *JI7*; yellow - *A. m. striatum;* grey - undetermined origin. The genotype of *ROS* and *EL* is given on the left and right, respectively. Individuals are numbered according to the attached file "*mapping_recombinants.xlsx*". Key markers are indicated by vertical lines and numbered as in Table 2.1. The location of the *ROS1-3* genes is indicated, with vertical lines denoting exons and horizontal lines denoting introns (notice a big intron in *ROS2*, which is 16.4kb long).

(Figure 4.9 continued)

A) Recombinants obtained by the genotypic screen strategy in Figure 4.6. Individual no. 17 carries a dominant allele of *ROS*, thus derived from *JI7* and a dominant allele of *EL*, thus derived from *striatum*. Therefore, its recombination point after marker 21 determines the leftmost border where *EL* is located. Individual no. 18 also carries the *ROS*^{7} allele, but it still carries the parental *el*^{7} allele. Therefore, its recombination point at marker 35 determines the rightmost border where *EL* is located. The two markers define an interval of 80kb.

B) Recombinant obtained by the phenotypic screen strategy in Figure 4.6. This individual (no. 50) defines a recombination point between *ROS* and *EL* further then that defined by individual no. 17 in A). Therefore the interval containing *EL* is reduced to 47kb. I note that the apparent absence of a "grey" portion of undetermined genotype is because the recombination point was between two markers separated by only 196bp (markers 27 and 28 in Table 2.1).

To further narrow the *EL* interval, a screening strategy to find recombinants by phenotype was used (phenotypic screening in Figure 4.6). In this experiment, heterozygous *ROS el* / *ros EL* individuals were crossed to the double-recessive *rosea*^{dorsea} mutant (*ros*^{d} *el*^{d} / *ros*^{d} *el*^{d}). The haplotypes of the heterozygote parent originated from *A. m. striatum* (*ros*^{s} *EL*^{s}) and from either *JI7* or *A. m. pseudomajus*, which are indistinguishable by phenotype (*ROS*$^{7/p}$ *el*$^{7/p}$). In the F1 of these test-crosses, the two haplotypes from the heterozygote parent segregate resulting in magenta (*ROS*$^{7/p}$ *el*$^{7/p}$ / *ros*^{d} *el*^{d}) or non-magenta (*ros*^{s} *EL*^{s} / *ros*^{d} *el*^{d}) progeny. These should be the commonest phenotypes observed in the F1 plants. However, if there is recombination between the two haplotypes in the heterozygote *ROS*$^{7/p}$ *el*$^{7/p}$ / *ros*^{s} *EL*^{s} parent, this may generate a haplotype of the type *ROS*$^{7/p}$ *EL*^{s}. When combined with the *rosea*^{dorsea} haplotype, in the F1, individuals carrying this recombinant haplotype will be detectable as having an Eluta phenotype (*ROS*$^{7/p}$ *EL*^{s} / *ros*^{d} *el*^{d}). This is because the dominant allele of *ROS*, which allows the production of pigment, becomes coupled with the semi-dominant allele of *EL*, which restricts that pigment

(Figure 4.6). The caveat of this approach is that an individual carrying the other recombinant haplotype (*ros*^{s} *el*$^{7/p}$ / *ros*^{d} *el*^{d}) is indistinguishable from the parental *striatum* haplotype that is also found in sib F1 plants (*ros*^{s} *EL*^{s} / *ros*^{d} *el*^{d}). Therefore, recombination rates between the two loci in this experiment are underestimated by half. However, the phenotypic screen has the advantage of allowing sampling a higher number of meiotic events, which would be impractical by the genotyping screening described above.

One batch of this screening was conducted between 2004 and 2009 by Annabel Whibley and Lucy Copsey (5248 plants) and a second batch screened by me in 2012 (5011 plants). In total there were 36 individuals with an Eluta phenotype (putative *ROS EL* recombinants) and 4 individuals with a pale but homogeneous magenta phenotype (these are discussed in section 4.2.2.2). To identify recombination points associated with *EL,* the 36 Eluta individuals were genotyped with markers linked to *ROS1*. From the 36 putative *ROS EL* recombinants, 10 surprisingly carried all three parental alleles at each of the genotyped markers (a *striatum*, a *rosea*^{dorsea} and a *JI7/pseudomajus* alleles). These individuals were deemed to have a chromosomal abnormality (discussed in section 4.2.2.3), and were thus ignored for mapping purposes. Therefore, this population produced 26 putative *ROS EL* recombinants.

The 26 recombinants provided a new estimate of the map distance between *ROS* and *EL* of $\frac{26}{10261} \times 100 \times 2 = 0.5\, cM$ (the multiplication by 2 is to correct for the fact that only half of the recombinants are detected). This figure is between the two previous estimates of 0.3 cM and 0.66 cM. The furthest recombination point from *ROS1* in these recombinants was 144.448 kb downstream of it. This reduces the previously determined map interval from 80kb to 47kb (Figure 4.9B). As this phenotypic screen depends on recombination events between *ROS* and *EL* (since recombination points beyond *EL* do not result in changes of phenotype), tightening the right border of the *EL* interval was not possible by this method.

The putative recombinant found in the *pseudomajus* x *striatum* F2 (individual E277-95 in Figure 4.5) was genotyped with a marker within the determined *EL* interval (marker 34, Table 2.1) and indeed confirmed to be a recombinant. This individual was homozygous for the *ROS1* markers ($ROS1^{P}$ / $ROS1^{p}$) but heterozygous for the marker in the *EL* interval (el^{p} / EL^{s}), thus confirming it has a $\underline{ROS^{p}\ EL^{s}}$ haplotype giving its Eluta phenotype.

4.2.2.2 Mapping recombination points in ROS

Although the *ROS* locus has been fairly well characterized (Schwinn *et al.* 2006), the existence of several *A. majus* alleles conferring different phenotypes (e.g. *Jl7*, $rosea^{dorsea}$ and $rosea^{colorata}$; Figure 4.1) suggests that different regions in this locus may control different aspects of the phenotype. Therefore, the recombinants primarily used to map *EL* were also investigated with regards to changes associated with the magenta phenotype conferred by *ROS*.

i) Mapping an interval associated with strong magenta pigment

The recombinants from the genotypic screen were used to define an interval associated with the production of magenta pigment by the ROS^{7} allele. A recombination point 74.720 kb upstream of *ROS1* defined the left border of such an interval: an individual that carried a *JI7* marker upstream of *ROS1* was still genetically determined to be *ros*^{s} (Figure 4.10A). On the other hand, a recombination point 93.341 kb downstream of *ROS1* defined the right border of the magenta interval (Figure 4.10A). This interval includes the *ROS1* as well as the *ROS2* and *ROS3* genes. The recombinants from the phenotypic screen further tighten this interval. These all carry a dominant allele of *ROS*, since all of them produce pigment (although being restricted by *EL*^{s}). From these recombinants, the closest recombination point to *ROS1* was 22.902 kb downstream of it, narrowing the previous interval to 100kb (Figure 4.10B). The recombinant individual defining this limit had a recombination between a marker in the third exon of *ROS1* and a marker in the third exon of *ROS2*. Therefore, it is undetermined if the other two exons of *ROS2* are from *pseudomajus* or *striatum* (notice they lie on grey area in Figure 4.10B), which does not allow excluding this gene as being necessary for the full magenta pigment. However, *ROS3* falls outside of the mapped interval, excluding it as necessary for the magenta phenotype of the flowers.

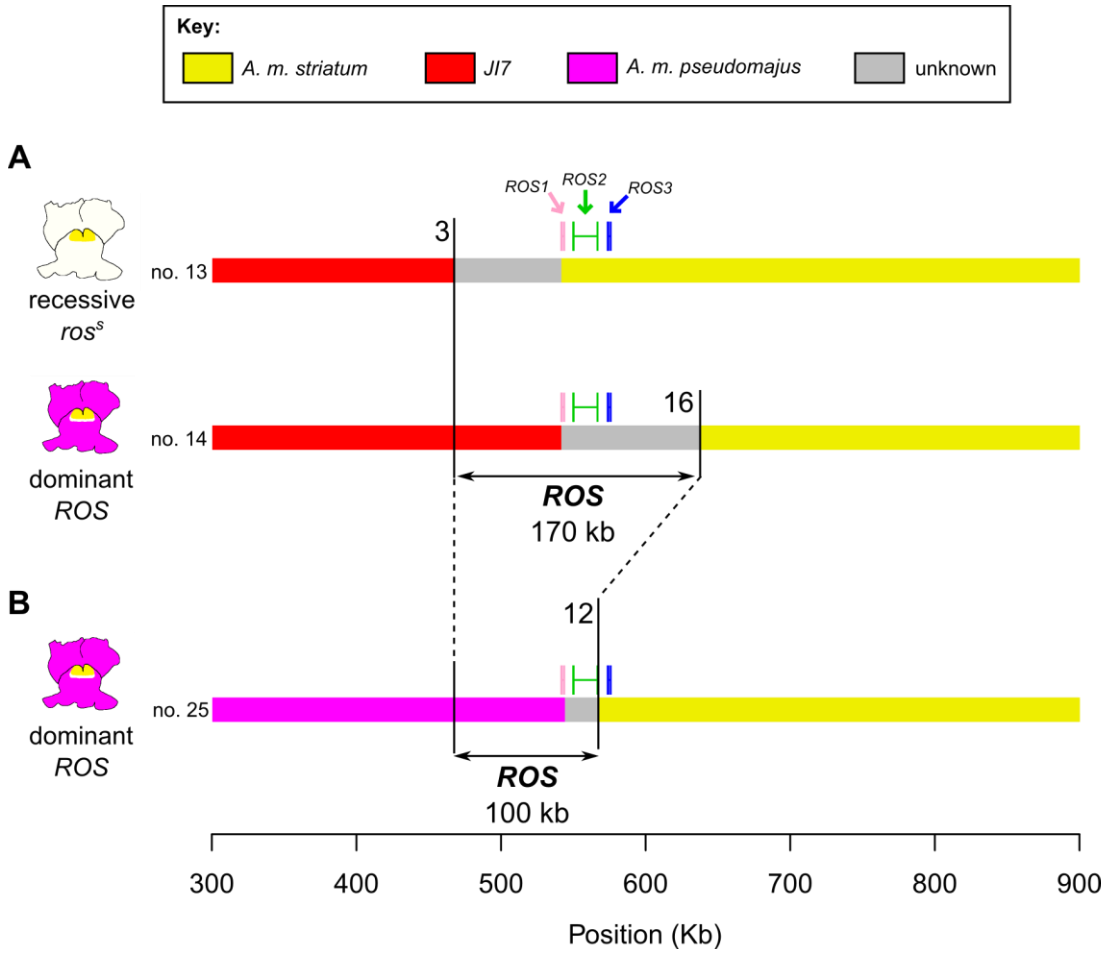


Figure 4.10 – Mapping recombination points flanking an interval associated with the production of strong magenta pigment.

Haplotypes are represented as in Figure 4.9, with the addition that *pseudomajus*-derived haplotypes are represented in magenta (to distinguish it from the *JI7* haplotype, in red).

A) Recombinants from the genotypic screen (Figure 4.6). Individual no. 13 has a recombination point upstream of *ROS1* and carries a recessive allele *ros*^{s}, thus defining a leftmost limit where this locus is located. Conversely, individual no. 14 has a dominant allele of *ROS*^{7} and a recombination downstream of *ROS3*, which defines the rightmost position of this locus. This interval spans 170kb and includes all three *ROS1-3* genes.

B) Recombinant from the phenotypic screen (Figure 4.6). This individual (number 25) defines a recombination point in the third exon of *ROS2* that narrows the interval defined by individual no. 14 in A). This reduces the interval to 100kb, which excludes the *ROS3* gene as necessary for producing pigment.

ii) The *A. m. striatum ros*^{s} allele confers pale pigmentation, which is restricted by *EL*^{s}

Although the *A. m. striatum ros*^{s} allele is unable to complement the *rosea*^{dorsea} mutation (Figure 4.3), it does not carry a knock-out mutation in *ROS*, since some pigment is still observed in the dorsal region of the petals in homozygous *ros*^{s} *EL*^{s} / *ros*^{s} *EL*^{s} individuals (Figure 4.11). The production of magenta pigment by the *A. m. striatum ros*^{s} allele is further clarified in homozygous *ros*^{s} *el*^{p}/*ros*^{s} *el*^{p} individuals: there is a very weak pigmentation in these flowers that is spread over the outer lobes of the flower, in an almost mirror-image of the typical Eluta phenotype (Figure 4.11). This phenotype resembles the phenotype of the *rosea*^{colorata} line (Figure 4.1C), which has a non-functional *ROS1* allele but a functional allele of *ROS2*, which is expressed in its flowers. Therefore, a hypothesis is that the pigment observed in *A. m. striatum* plants is due to the contribution of a functional *ROS2* allele in this subspecies.

The difference between the phenotype conferred by *ros*^{s} *EL*^{s} and *ros*^{s} *el*^{p} haplotypes shows that the *ros*^{s} allele is not only able to produce low amounts of pigment, but it also responds to the pigment restriction determined by the dominant *EL*^{s} allele. Therefore, the non-magenta phenotype of *A. m. striatum* plants (*ros*^{s} *EL*^{s} / *ros*^{s} *EL*^{s}) is not only due to a weaker, recessive, allele of *ROS*, but also due to the effect of *EL*, which reduces the pigmentation in the flowers conferred by the former.

Despite the visible pigmentation in homozygous *ros*^{s} *el*^{p}/*ros*^{s} *el*^{p} individuals, this recombinant haplotype was not visually detected in the phenotypic screen (except for unusual exceptions detailed in the next section). This is likely due to the fact that, in the phenotypic screen, all screened plants are heterozygous with *rosea*^{dorsea}, and therefore *ros*^{s} *el*^{p}/*ros*^{dor} *el*^{dor} individuals might be undetectable if the *ros*^{s} is not fully dominant over the *ros*^{dor} allele.

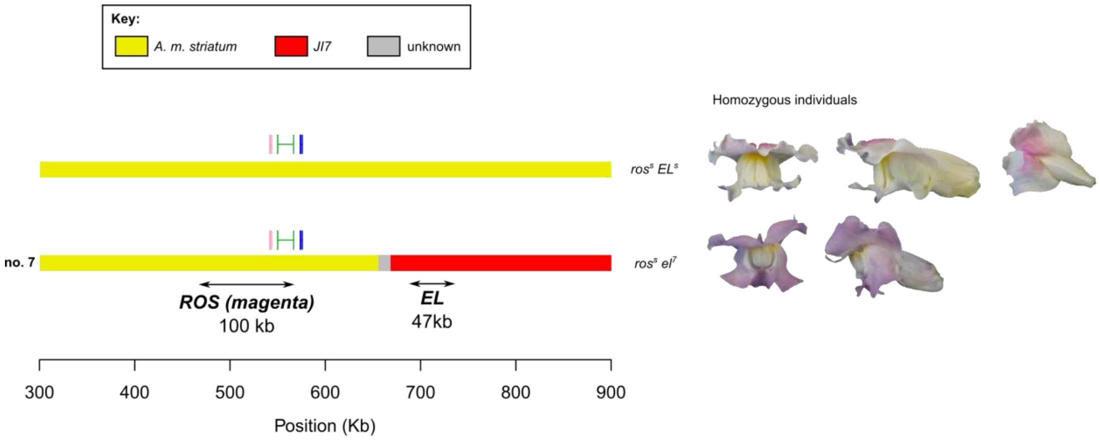


Figure 4.11 – Production of pigment by the *A. m. striatum ros*^{s} allele.
Individuals homozygous *ros*^{s} *EL*^{s}/ *ros*^{s} *EL*^{s} are able to produce some pigment in the dorsal region of the lobes (top). This is likely due to some activity of the *ROS* locus from *A. m. striatum*. In a recombinant where this allele is coupled with a recessive allele of *el*^{7} (*ros*^{s} *el*^{7}/ *ros*^{s} *el*^{7}), a very pale pigmentation becomes apparent in the lobes of the flowers (bottom). The intervals corresponding to *ROS* and *EL* are indicated (from Figure 4.9 and Figure 4.10).

iii) Production of pigment by elements tightly linked to *ROS1*

The involvement of elements tightly linked to *ROS1* and *ROS2* was elucidated by the analysis of recombinant plants with an unusual flower pigmentation detected in the recombinant screens. These individuals had flowers paler than *JI7*, but darker than the phenotype of *ros*^{s} *el*^{p}/*ros*^{s} *el*^{p} individuals described above (I will call this phenotype “rosy”; Figure 4.12). These rosy individuals are generally characterized by a pale tube (with pigmentation stronger at its base) and a pale magenta pigmentation in the lobes of the flowers. Whereas the Eluta phenotype confers stronger pigmentation in the central part of the lobes, the pigmentation in the rosy phenotype is overall paler on the lobes, not being stronger magenta in the centre, like Eluta (Figure 4.12).

In terms of phenotype, the *ros*^{rosy} allele is not fully dominant over the *ros*^{dor} allele, since the pigmentation of heterozygous *ros*^{rosy} *el*^{p}/*ros*^{dor} *el*^{dor} is paler than the homozygous recombinant *ros*^{rosy} *el*^{p}/*ros*^{rosy} *el*^{p} (Figure 4.12). However, *ros*^{rosy} seems to be fully recessive to the *ROS*^{7} allele, since heterozygous *ros*^{rosy} *el*^{p}/*ROS*^{7} *el*^{7} have a strong magenta phenotype (Figure 4.12).

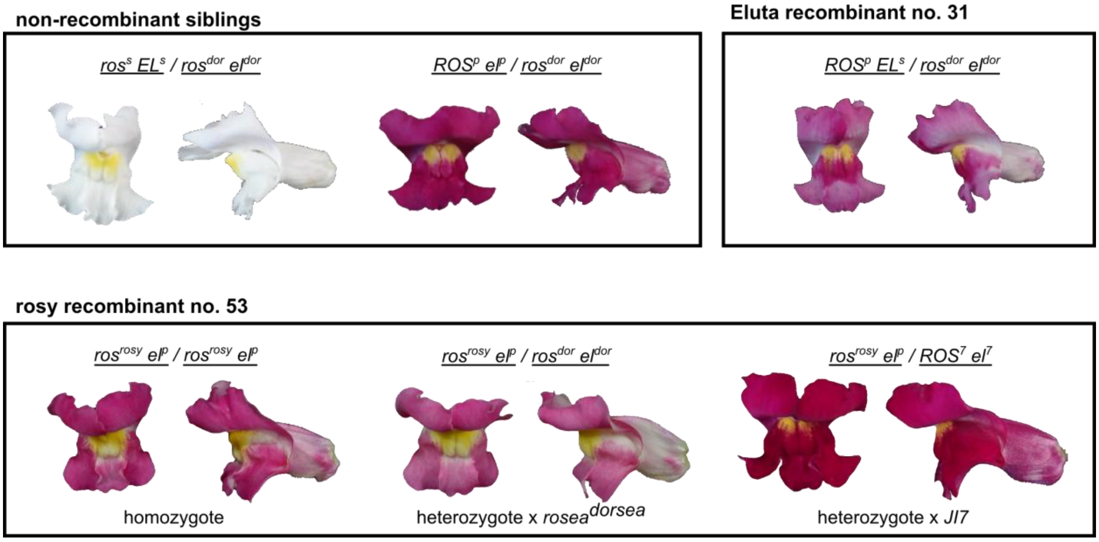

non-recombinant siblings
ros^{s} EL^{s} / ros^{dor} el^{dor}
ROS^{p} el^{p} / ros^{dor} el^{dor}
Eluta recombinant no. 31
ROS^{p} EL^{s} / ros^{dor} el^{dor}
rosy recombinant no. 53
ros^{rosy} el^{p} / ros^{rosy} el^{p}
ros^{rosy} el^{p} / ros^{dor} el^{dor}
ros^{rosy} el^{p} / ROS^{7} el^{7}
homozygote
heterozygote x $rosea^{dorsea}$
heterozygote x JI7

Figure 4.12 – Phenotypes of recombinants and non-recombinant siblings from the phenotypic screen.

Photographs are shown for non-recombinants (magenta and non-magenta phenotypes), a recombinant between *ROS* and *EL* (Eluta phenotype) and a recombinant within the *ROS* locus (rosy phenotype). Notice the difference between the Eluta phenotype (stronger pigmentation in the central part of the lobes) and the rosy phenotype (weaker pigmentation in the lobes, not pronounced in the centre). The phenotype of a rosy haplotype heterozygous with the *rosea*^{dorsea} haplotype has the same pattern of pigmentation as the homozygote, but less intense. This suggests that the *ros*^{rosy} allele is not fully dominant over the *ros*^{dor} allele. Conversely, a heterozygote with the *JI7* haplotype is magenta, suggesting *ros*^{rosy} is recessive to *ROS*^{7}.

One of these paler rosy individuals was found in the genotypic screen, whereas four others were found in the phenotypic screen. All 5 individuals carried a *ros1*^{s} marker, and had recombination points between this gene and the third exon of *ROS2* (Figure 4.13). This suggests that this component of the colour is conferred by *pseudomajus* or *JI7* elements downstream of *ROS1*. The interval defined for this element is 126 kb, which includes the *ROS2* and *ROS3* genes (Figure 4.13). Despite this large map interval, the fact that all 5 of these individuals had recombination points between markers in *ROS1* and the third exon of *ROS2* suggests that the elements contributing to the darker tinge of magenta in rosy flowers, are tightly linked to *ROS2* (otherwise, by chance, rosy individuals should have been detected with recombination points downstream of this gene).

Even though the pattern of pigmentation was similar between the rosy recombinants, there was visible variation in the intensity of magenta pigment each of them had. This could be due to particular differences in the exact recombination point between each haplotype (which remains to be narrowed down with the addition of more markers).

In summary, three main components of the magenta phenotype were mapped by the analysis of recombinants: a 46kb interval containing *EL*; a component of *ROS* associated with the strong magenta phenotype (likely involving *ROS1*); and another component of *ROS* associated with the production of small amounts of pigment (likely involving *ROS2* or regulatory elements of *ROS1*). This reveals that the *ROS-EL* region includes a tight cluster of loci that, together, determine the final magenta pigmentation of the flowers.

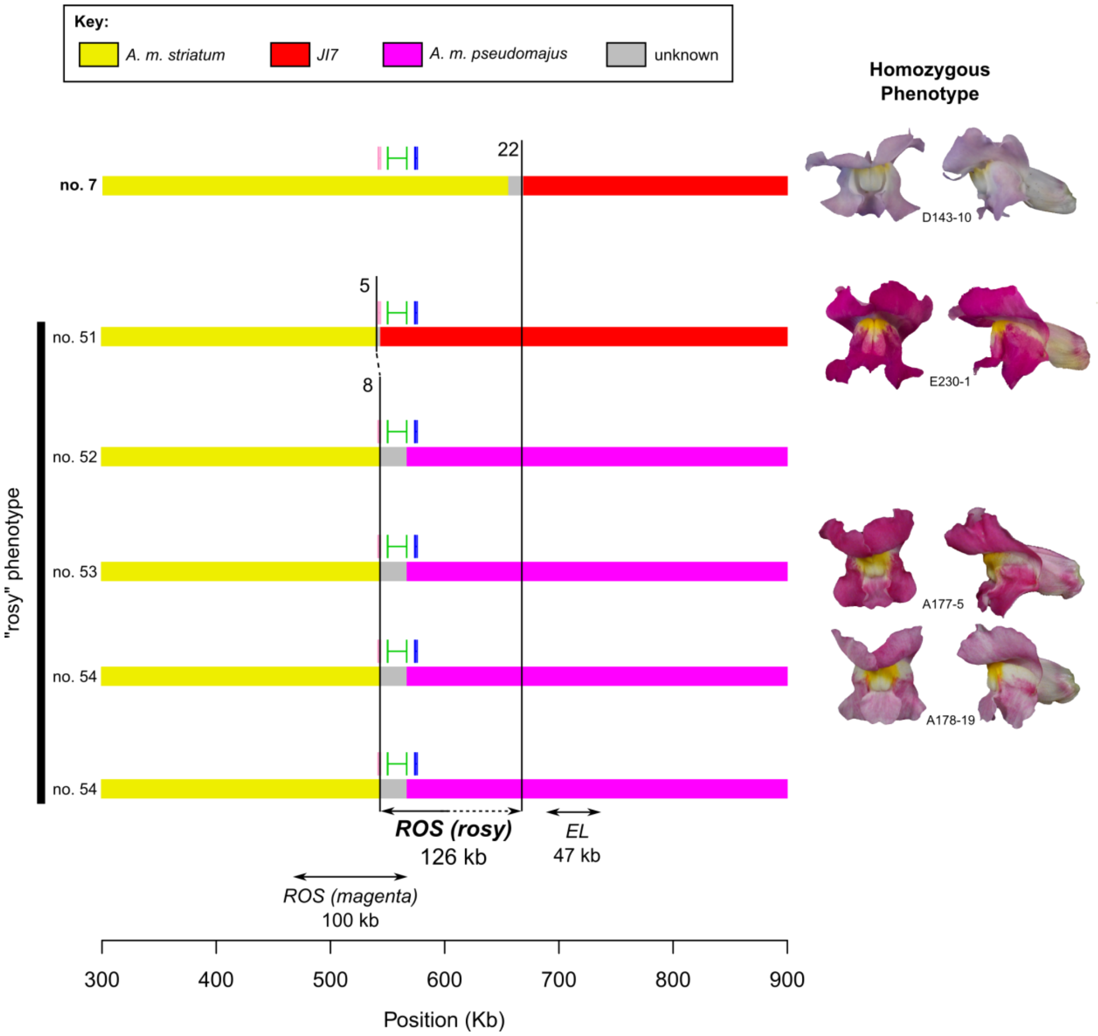


Figure 4.13 – Mapping recombination points associated with "rosy" phenotype. Haplotypes represented as in Figure 4.9 and Figure 4.10. A photograph of a flower from individuals homozygous for each haplotype is shown, if available (individual numbers given for reference). Individuals with pale magenta phenotype ("rosy") have recombination points just downstream of *ROS1*. They define the left border of an interval containing elements that contribute to the partial production of magenta pigment in the flowers. A recombination point downstream of *ROS2* no longer produces the rosy phenotype (recombinant no. 7 discussed in Figure 4.11), defining a right border for the "rosy" interval. Although the mapped interval is large, the fact that all 5 rosy individuals had recombination points near *ROS1* suggest this is due to an element close to this gene (the partially dashed arrow denotes this), possibly related to *ROS2*.

4.2.2.3 Characterization of tri-allelic genotypes

Some of the putative recombinants obtained from the phenotypic screen had three alleles for some of the markers used to map recombination points. In any given test-cross three haplotypes are in the parents: two haplotypes from the heterozygote parent ($\underline{ROS^{7/p}\ el^{7/p}}$ / $\underline{ros^{s}\ EL^{s}}$) and one haplotype from the *rosea*^{dorsea} parent ($\underline{ros^{dor}\ el^{dor}}$ / $\underline{ros^{dor}\ el^{dor}}$). We usually expect that any F1 progeny from such crosses should have only two of these haplotypes. However, 10/10261 individuals from the phenotypic screen contained three haplotypes, based on 5 markers (example in Figure 4.14).

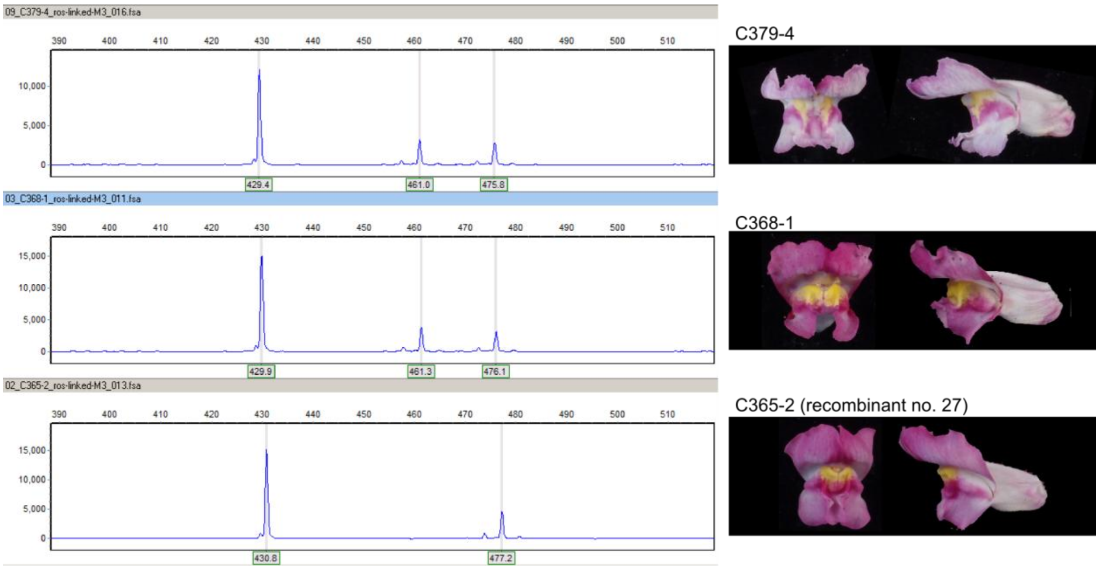


Figure 4.14 - Example of individuals with tri-allelic genotypes.

A snapshot of a microsatellite electropherogram is shown for two individuals with three alleles (top two) and one individual with a normal diploid genotype (bottom). All individuals are derived from a phenotypic screen for *ROS-EL* recombinants. Photos of each individual are shown on the right (the ID of each is given for reference).

The consistency between the 5 markers excluded the possibility of the triple haplotypes being due to a spurious genotyping error. However, they could be due to contamination of the DNA samples, resulting in DNA of mixed origin. I excluded this by repeating the DNA extraction from a single leaf and re-genotyping the individuals. The result did not change: three haplotypes were present for all informative markers. These markers span a region of ~400 kb, which suggests that these individuals have three copies of at least this portion of the genome.

Three hypotheses were considered to explain this result. One is that this region is tandemly duplicated due, for example, to an unequal crossing-over between the parental haplotypes. Another hypothesis is that these individuals have a trisomy of the *ROS* chromosome due, for example, to a failure in the chromosome separation during meiosis. Finally, the region could be duplicated elsewhere in the genome due to a translocation of a portion of the *ROS* chromosome to a non-homologous chromosome. The first hypothesis (tandem duplication) can be distinguished from the latter two (trisomy or non-homologous translocation) by looking at segregation of markers in self-progeny F2s from these individuals. A tandem duplication should result in the non-independent segregation of some alleles, since the duplicated regions are linked in the same chromosome arm. Conversely, a trisomy or translocation should result in the independent segregation of the marker alleles.

I analysed 96 self-progeny from one of these individuals ($\underline{ROS^{p}\ el^{p}}$ / $\underline{ros^{s}\ EL^{s}}$ / $\underline{ros^{dor}\ el^{dor}}$) and looked at the segregation of the 5 polymorphic markers that distinguished the three parental alleles. The three alleles in all markers segregated independently (Table 4.1), excluding a tandem duplication as an explanation for this result.

Table 4.1 – Segregation of alleles in self-progeny of an individual with three copies of the *ROS-EL* region.

The genotype is given for a marker in *ROS1*, but 4 other markers linked to it agreed with this genotype. The observed and expected (for a trisomy) numbers are given for each genotype. The family used in this experiment was E350.

***ROS1* Genotype**	**Phenotype**	**Observed**	**Expected**
ROS^p / ROS^p	Magenta	6	3
ros^s / ros^s	Non-magenta	13	3
ros^d / ros^d	Non-magenta	6	3
ROS^p / ros^s	Eluta	22	19
ROS^p / ros^d	Magenta	22	19
ros^s / ros^d	Non-magenta	26	19
ROS^p / ros^s / ros^d	Did not flower	1	30
total		96	

The distinction between a translocation and a trisomy was possible due to a crucial difference between these two hypotheses. For a translocation, there are two "real" *ROS* chromosomes and one "recipient" chromosome where a *ROS* haplotype was translocated to. Any gamete with the "recipient" chromosome will always carry one of the "real" *ROS* chromosomes. Therefore, one of the *ROS* haplotypes (the translocated one) should never occur in a homozygous state, under this hypothesis. In opposition, in a trisomy each of the three chromosome copies can segregate independently, and therefore all haplotypes may potentially occur in a homozygous state. In the 96 self-progeny analysed, homozygous individuals were found for all three haplotypes (Table 4.1), supporting in favour of a trisomy in the parent that gave origin to this family.

The genotype ratio in the analysed family does not fit with the expected independent segregation of the three alleles in the trisomic parent (Table 4.1), likely due to severe segregation distortions in trisomic individuals. Only one individual in the 96 self-progeny was recovered with three copies of the haplotype, and it had a phenotype distinct from its sibs: it had rounder leaves, shortened height, a fragile structure (thin stems) and it did not produce flowers. Further confirmation of the trisomy could be aided by analysing the karyotype of these individuals and looking to see if there is an extra chromosome.

Whichever chromosomal aberration explains this result, it is a relatively rare event occurring at a frequency of 0.1%. The reason it was picked in 10 individuals is that screening for Eluta phenotype in the test-crosses identifies recombinants as described in section 4.2.2.1, as well as these chromosomal aberrations which bring together a dominant allele of *ROS* from the *pseudomajus/JI7* haplotype with a dominant allele of *EL* from the *striatum* haplotype.

These rare occurrences could confuse future screens of this kind, but they are easily resolved by genotyping individuals with several polymorphic markers that distinguish all haplotypes involved. As these cases do not contribute to the mapping of the colour loci, their characterization was not carried further.

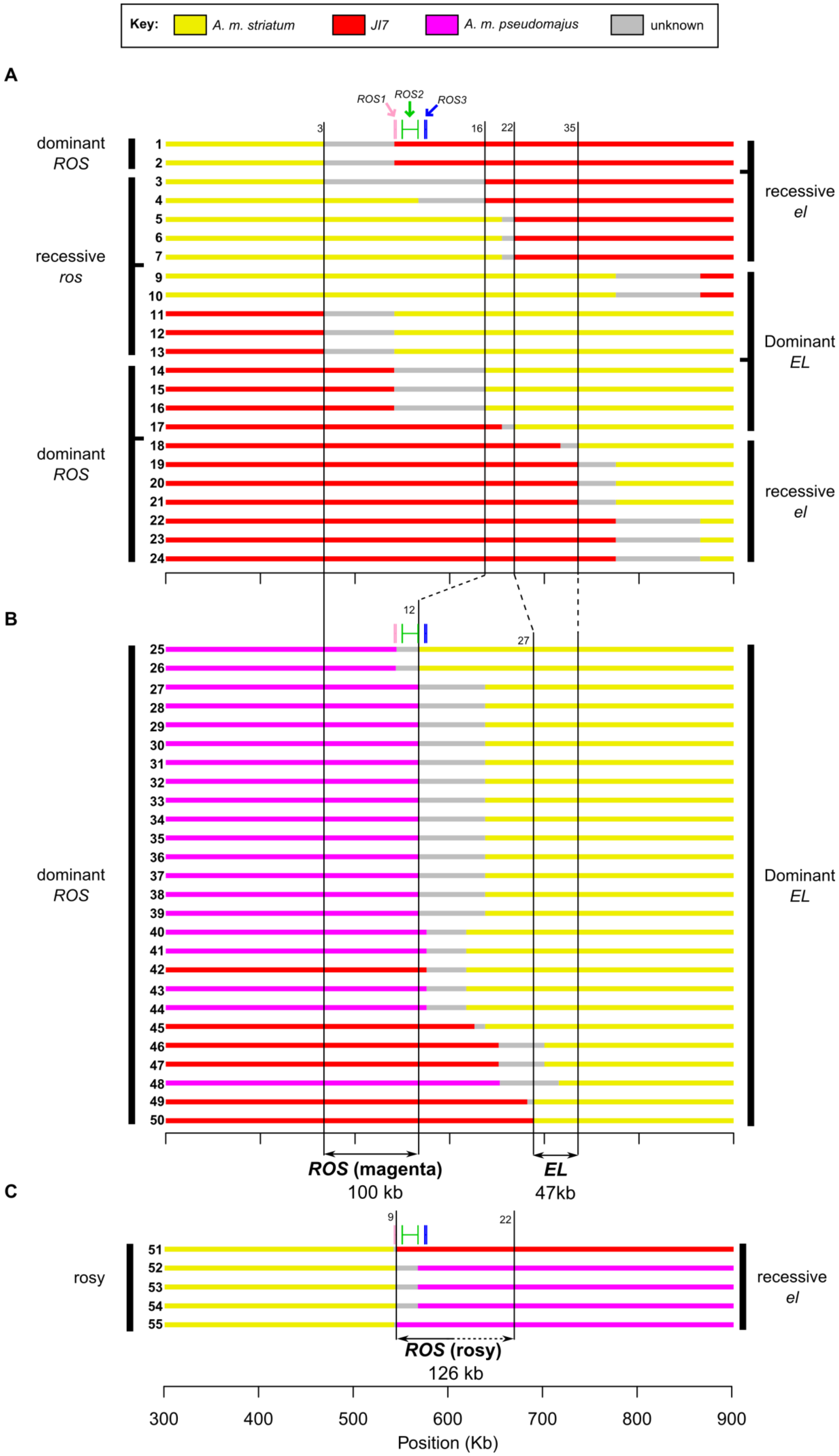

Key:
A. m. striatum
JI7
A. m. pseudomajus
unknown
A
ROS1
ROS2
ROS3
3
16
22
35
dominant ROS
recessive ros
dominant ROS
recessive el
Dominant EL
recessive el
1
2
3
4
5
6
7
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
B
12
27
dominant ROS
Dominant EL
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
ROS (magenta)
100 kb
EL
47kb
C
9
22
rosy
recessive el
51
52
53
54
55
ROS (rosy)
126 kb
300
400
500
600
700
800
900
Position (Kb)

Figure 4.15 (previous page) – Full panel of recombinants used to map recombination points around *ROS*.

Horizontal coloured lines represent a haplotype determined from molecular markers used to genotype the plants. Magenta – *A. m. pseudomajus;* red – *A. majus JI7*; yellow – *A. m. striatum*; grey – unknown. Each individual haplotype is numbered as in the attached file "*mapping_recombinants.xlsx*". Key markers flanking mapped loci are indicated as vertical lines and numbered as in Table 2.1. The *ROS* and *EL* genotypes are indicated on the left and right, respectively. The location of the *ROS1, ROS2* and *ROS3* genes is indicated, with vertical lines denoting exons and horizontal lines denoting introns (notice a big intron in *ROS2*, which is 16.4kb long). The mapped intervals correspond to *EL*, and two regions of *ROS* (producing the stronger magenta and the weaker rosy). These are detailed in Figure 4.9, Figure 4.10 and Figure 4.13, respectively.

A) Recombinants from genotypic screen.

B) Recombinants from phenotypic screen.

C) Recombinants with rosy phenotype.

4.3 Discussion

4.3.1 Mapping *EL*

Different strategies have been used to map *ROS* and *EL*. The earliest evidence of recombination between these loci comes from the early 20th century work done by Baur (Baur 1911, 1912). This work was being carried out before the concept of genetic linkage and the chromosome theory were established in the field of genetics. Therefore, Baur's experiments were a pioneering attempt to understand the phenomenon of non-independent segregation of loci, which had been previously reported by Bateson, Punnett and Saunders in sweet peas (Bateson *et al.* 1905). Through a series of crosses between several lines of *A. majus*, Baur demonstrated that *ROS* and *EL* (in his notation they were called factors "F" and "G", respectively) did not segregate independently. However, he extended his experiments to show that the linkage (or "coupling", as it was called) was not absolute, and that some reversions of phenotype could be observed (which can now be interpreted as recombination between the loci). I re-visited his results and collected the numbers of segregating progeny from his crosses, which are schematized in Figure 4.16. His strategy allows recombinants with magenta phenotype to be identified from the self-progeny of a heterozygous *ROS EL* / *ros*^{dor} *el*^{dor} plant. The magenta individuals are *ROS el*^{dor}, which are equivalent to *JI7* and *pseudomajus* haplotypes. Although this is the first evidence of recombination between the two loci, Baur's results are hard to explain. According to his segregation ratios, the distance between *ROS* and *EL* is around 24cM (explained in Figure 4.16), a distance much higher than the 0.5cM reported in this chapter. In fact, Baur's strategy for finding recombinants was attempted by Schwinn et al. (2006), but they failed to find any recombinants (although the size of the progeny is not reported). I cannot say why Baur's results are incongruent with more recent mapping efforts, but it could be related to some specific characteristics of the particular *A. majus* lines he used in his crosses (which are unknown to me).

More recent evidence of recombination between *ROS* and *EL* comes from Whibley et al. (2006), who found two recombinants out of 1300 progeny from a test-cross (phenotypic screen in Figure 4.6), giving a map distance of $\frac{2}{1300} \times 2 \times 100 = 0.3cM$ between *ROS* and *EL*. This figure agrees with the results in this chapter. I obtained three estimates from three independent experiments: the *pseudomajus* x *striatum* F2 (section 4.2); the *JI7* x *striatum* F2 (genotypic screen in Figure 4.6); and the test-crosses (phenotypic screen in Figure 4.6). The estimates were 0.3cM, 0.66cM and 0.5cM, respectively. The different estimates are likely due to the sample size being different between them, respectively 322, 1370 and 10261 meiotic products analysed. I consider the most accurate estimate to be the one with highest sample size, thus I consider *ROS-EL* as being 0.5cM apart.

Using several polymorphic molecular markers, I was able to identify recombination points surrounding the *EL* locus, identifying a 47kb region ~144kb downstream of *ROS* that contains this locus. This provides a relatively narrow interval where candidate genes for *EL* can be screened (discussed in chapter 5).

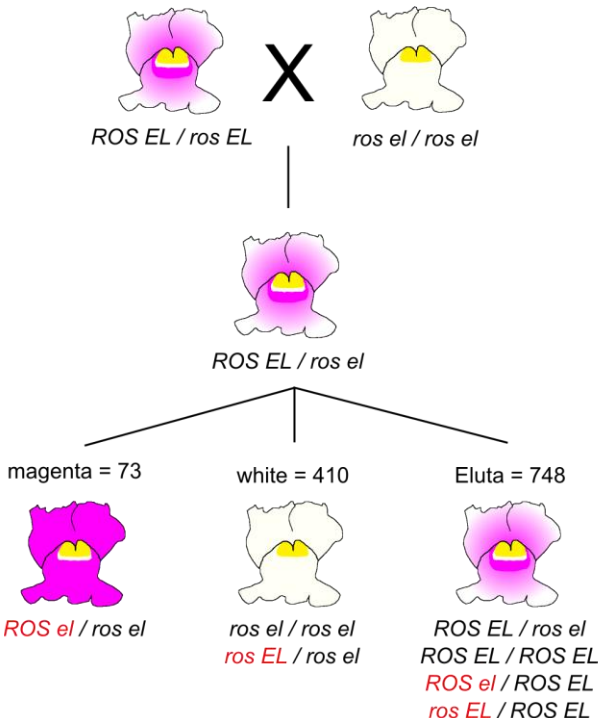


Figure 4.16 – Results from Baur's segregation experiments with *ROS-EL*.
His crossing strategy is schematized and the numbers of each phenotype obtained in the segregating population are shown (numbers from Table I in Baur 1912). The possible genotypes are given for each phenotype, with the recombinant haplotypes highlighted in red. Only *ROS el* recombinants producing magenta flowers can be distinguished by phenotype. The frequency of magenta plants is 73/1231 = 6%. Because *ROS el* recombinant haplotypes are only distinguished when combined with the parental *ros el* haplotype, their frequency can be approximated by doubling this value, i.e. 12%. The other type of recombinant (*ros EL*) is always indistinguishable by phenotype, which means that an estimate of the total recombination frequency should double the last value, giving an approximate distance between *ROS* and *EL* of 24cM.

4.3.2 Mapping flower colour phenotypes within the *ROS* locus

The recombinants obtained for mapping *EL* were further used to characterize phenotypes associated with the *ROS* locus. Although this locus has been cloned, the occurrence of at least three tandemly duplicated *ROS*-like genes suggests that they may regulate different aspects of the colour. I have identified an interval of 100kb associated with strong production of magenta pigment, which contains the *ROS1* and *ROS2* genes. This interval excludes *ROS3*, suggesting this gene is not involved in producing pigment in the flowers of *JI7* and *A. m. pseudomajus*. The work of Schwinn et al. (2006) suggests that *ROS1* is sufficient for the production of pigment in *JI7* magenta plants and my mapped interval is compatible with this hypothesis.

It should be noted that *JI7* and *A. m. pseudomajus* flowers differ in the intensity of the magenta pigment (e.g. in Figure 4.3). Generally, *A. m. pseudomajus* flowers have less intense pigmentation than *JI7*. Moreover, in the F2 progeny of *A. m. pseudomajus* x *A. m. striatum* there is variation in the intensity of magenta, both within the Eluta and magenta phenotype classes (see variation in scores in Figure 4.5). This suggests that modifier loci may be involved in some of the differences in pigmentation, which may not be controlled by *ROS*. My mapping experiments did not account for this variation, as they were all focused on *ROS-EL* and most recombinant screens occurred in crosses with *JI7* or *rosea*^{dorsea} lines, thus masking the contribution of unlinked modifier loci.

An unexpected phenotype appeared in the recombinant screens, characterized by a paler magenta pigment in the flowers. This phenotype, named “rosy”, was associated with recombination points between the third exon of *ROS1* and the third exon of *ROS2*. This suggests that elements downstream of *ROS1* can be involved in producing some pigment in the flowers. Indeed, the *rosea*^{colorata} mutant (Figure 4.1) has a similar phenotype to these plants. This mutant carries a loss-of-function *ROS1* allele but expresses *ROS2* at higher levels than either *JI7* or *rosea*^{dorsea} (Schwinn *et al.* 2006). This suggests that *ROS2* could be involved in producing some pigment, perhaps being somewhat redundant with *ROS1*. The mapped interval associated with the "rosy" phenotype is compatible with this result, suggesting that it could be due to a *ROS2*^{p} allele carried by these recombinants (Figure 4.13).

The most parsimonious explanation for these results is that *ROS1* and *ROS2* both contribute to the magenta pigment seen in *A. m. pseudomajus* flowers (Figure 4.17). In these flowers, the pigment is spread on the corolla due to the presence of a recessive *el*^{p} allele in the background. On the other hand, *A. m. striatum* has a partially functional *ROS* locus, but its effect is masked by the presence of a dominant allele of *EL*^{s} in the background (this is apparent in recombinants $\underline{ros1^{s}\ ros2^{s}\ el^{p}}$).

It still remains to clarify if the *ros1*^{s} allele carried by rosy recombinants is contributing to the phenotype, perhaps by interacting with *pseudomajus* elements downstream of it. Also, I cannot say which parts of the *ROS2* gene these recombinants carry. Currently, the marker used to genotype them was located on the third exon of *ROS2*, therefore I cannot assess the origin of the other two exons. This should be improved in the future by finding new markers within this region. It could be the case that some of the variation in the rosy phenotype is related to the particular location of the recombination point in each of them. For example, if multiple polymorphisms in this region interact (additively or epistatically) to produce the final colour of the flowers, it could be that the differences in phenotype intensity conferred by different rosy haplotypes is due to the particular combination of functional polymorphisms that each carry.

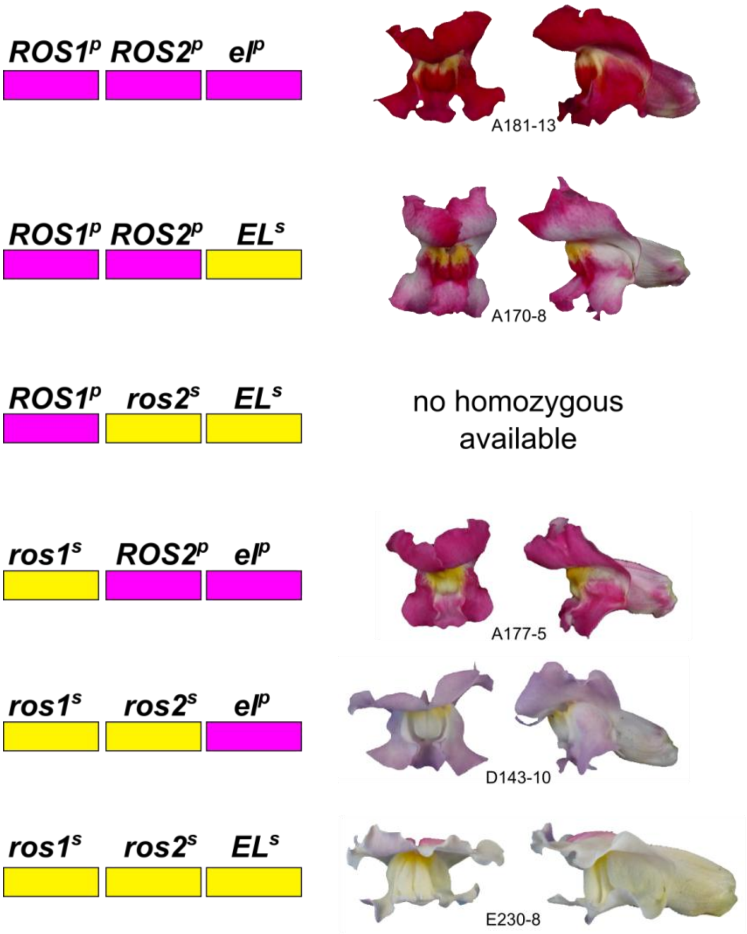


Figure 4.17 – Hypothesis to explain different phenotypes observed in recombinant analysis.
Six possible allelic arrangements are given for *ROS1*, *ROS2* and *EL*. Photographs of homozygous individuals with each haplotype are shown.

The *ros*^{rosy} allele seems to be haploinsufficient, that is, a single copy of this allele in *ros*^{rosy}/*ros*^{dor} is not enough to produce the phenotype seen in homozygotes *ros*^{rosy}/*ros*^{rosy} (Figure 4.12). I did not cross this recombinant haplotype with a *ros EL* haplotype, thus I cannot confirm that a dominant *EL* also restricts the pigment conferred by the *ros*^{rosy} allele. If this were the case, it would mean that *EL* is also able to interact with the elements that produce the pigment in rosy recombinants.

In summary, the dissection of phenotypic changes associated with the broadly defined *ROS* locus, reveal genetic interactions within this locus. The genetic analysis presented in this study reveals that multiple polymorphisms within this region may interact with each other, and this could be relevant in the context of the *A. m. pseudomajus* and *A. m. striatum* hybrid zone. Particularly, some of the variation in magenta intensity seen in hybrid zone individuals (and in F2 families, as discussed above) could also be related to different *ROS* alleles segregating in the populations. These may not only interact in hybrid individuals, producing darker or lighter phenotypes, but also with modifiers in the background (being more or less influenced by them).

The *ROS-EL* locus may not be unique to *Antirrhinum.* For example, in *Mimulus*, there is a likely homologous region to the *Antirrhinum ROS-EL*, named *pla1* locus (in scaffold 11 of *M. guttatus*; Cooley et al. 2011). This region contains several duplicated *MYB*-like genes (one of them found through homology with *ROS1*) and it co-segregates with several traits related to anthocyanin pigmentation (Lowry *et al.* 2012). The *pla1* locus is polymorphic in wild populations of *M. guttatus*, being both variable within populations as well as fixed between other populations. In the sister species, *M. aurantiacus* the homologous gene to *ROS1* (called *MaMYB2*) is also responsible for the magenta difference between two morphs that occur in nature. Further, this gene shows an allelic cline in a hybrid zone between the two morphs, suggesting it is under selection (Streisfeld *et al.* 2013).

In *Petunia*, the homologous gene to *ROS1* (called *AN2*) was also shown to be responsible for the phenotypic difference in flower colour between the species *P. integrifolia* and *P. axillaris* (Quattrocchio *et al.* 1999; Hoballah *et al.* 2007). In another pair of species, *P. axillaris* and *P. exserta*, a cluster of loci regulating several aspects of floral morphology (including flower colour) was identified as distinct between the two species (Hermann *et al.* 2013). Similarly, this cluster includes several *MYB*-like genes homologous to known pigmentation genes.

A good example of how multiple independent mutations influence different components of a selected trait, is the determination of coat colour in deer mice, *Peromyscus maniculatus* (Linnen *et al.* 2013). Different populations of this species have distinct coat colour phenotypes, adapted to particular environments. This difference in phenotype is due to several independent mutations within the *agouti* locus, which are associated with different components of the coat colour (e.g. brightness, ventral colour, tail stripes). In another species of this genus, oldfield mice (*P. polionotus*), the same *agouti* gene is involved in the coat colour difference between different populations (Steiner *et al.* 2007). However, a second locus, *Mc1r*, interacts epistatically with *agouti* by changing the pigmentation in cheek hairs only in certain genetic backgrounds (similarly to the effect of *EL* being only visible in when dominant *ROS* is in the background).

Generally, these examples illustrate that traits thought to be under selection (like flower colour) may often involve clusters of several loci (e.g. genes, cis-regulatory elements, point mutations, etc.), whose allelic combinations may be maintained by selection. Particularly, mutations that may be neutral in one background may not be so in another background. For example, if the *ROS1*^{p} allele is enough to produce the full magenta pigment in flowers, the contribution of *ROS2* to the phenotype is redundant and therefore mutations that affect its expression may be neutral in *A. m. pseudomajus*. However, in an *A. m. striatum* individual, activity of *ROS2* may not be neutral (if some pigment is produced, as seen in rosy recombinants), therefore mutations that reduce its activity may be positively selected in this background.

4.3.3 Relating mapped loci with divergence between *A. m. pseudomajus* and *A. m. striatum*

The genetically mapped intervals associated with different aspects of flower colour can be compared with the divergence between *A. m. pseudomajus* and *A. m. striatum* discussed in Chapter 3 (Figure 4.18). The three prominent *Fst* peaks emphasized in the *ROS* scaffold lie within the mapped intervals. In particular, the first peak co-localizes with *ROS1* and the first two exons of *ROS2*, suggesting divergence between *A. m. pseudomajus* and *A. m. striatum* in the locus (or loci) conferring the full magenta pigment of the flowers. The second peak co-localizes with the third exon of *ROS2* and, marginally, with *ROS3* (*ROS3* is at the edge of the peak). Finally, the furthest peak downstream of *ROS* co-localizes with the interval where *EL* was mapped, suggesting it is also divergent between the two populations in the *A. m. pseudomajus* and *A. m. striatum* hybrid zone.

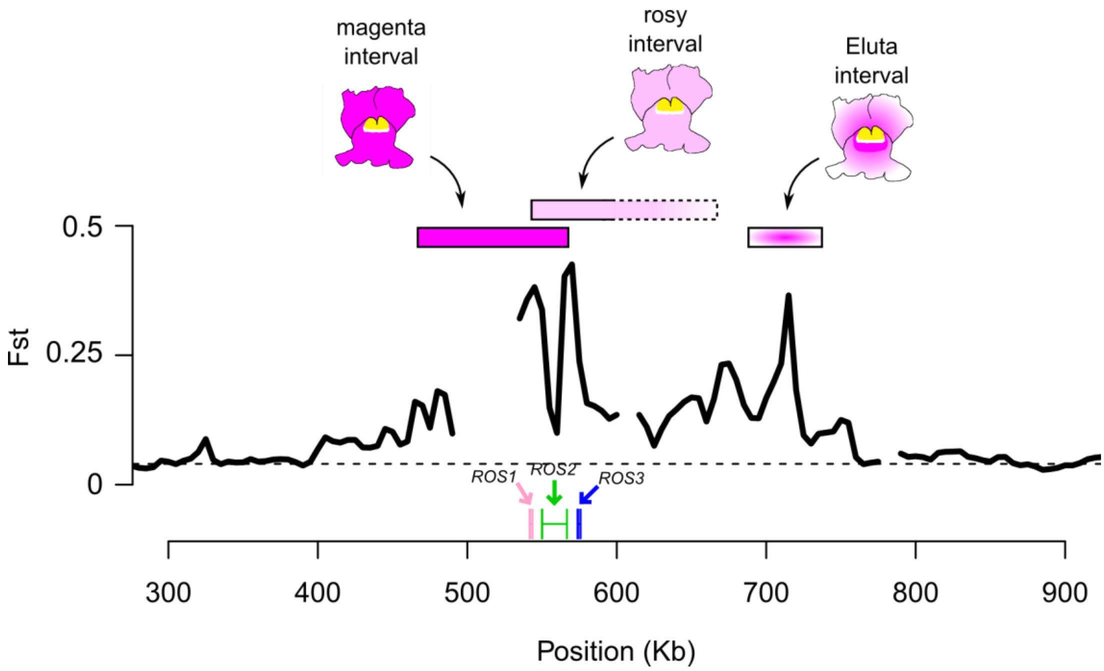


Figure 4.18 – Relating Fst with genetic intervals associated with flower colour loci. 10kb-window-averaged *Fst* across the *ROS* scaffold between samples from different sides of the *pseudomajus x striatum hybrid zone* (pools YP1 and MP4; Table 3.1). The horizontal dashed line is the genome-wide *Fst* median. The location of the *ROS1* to *ROS3* genes is indicated (vertical lines correspond to exons and horizontal lines to introns). The genetic intervals mapped in this chapter are indicated as coloured boxes above the plot.

Broadly speaking, there is a clear signal of divergence – high *Fst* – in both *ROS* (with its putative sub-components) and *EL*, whereas regions between them have lower divergence – near-average *Fst* – indicating that they are more similar between *pseudomajus* and *striatum*. Because both *ROS* and *EL* have similar levels of *Fst*, it could be that they are under similar selective pressures, which is further supported by their involvement in regulating the same trait (flower colour). In particular, the fact that there is genetic epistasis between them (for example, the dominant *EL* is only visible when a dominant *ROS* is present in the background) suggests that these loci might occur as co-adapted allelic combinations in each species: *pseudomajus* with *ROS el* and *striatum* with *ros EL*. If selection acts to maintain these combinations (even if there is gene flow, as in the hybrid zone), then the divergence between the two subspecies is expected to be similar in both loci, as seen with *Fst* (Figure 4.18).

In *Heliconius* butterflies, a very similar case to *Antirrhinum* is observed related to the control of wing colour polymorphisms. For example, a chromosome region associated with the control of red pigmentation in the wings is highly divergent between morphs of *H. melpomene* and between morphs of *H. erato* (Nadeau *et al.* 2012; Supple *et al.* 2013). The profile of divergence (*Fst*) across the region is heterogeneous, with at least two major sub-regions highly associated with wing-colour phenotypes. These two sub-regions co-localize with two genes, *optix* and *kinesin*, both associated with the production of red pigment in the wings (Pardo-Diaz & Jiggins 2014).

These parallels with *Antirrhinum* suggest that clusters of loci regulating divergent traits may often be responsible for the polymorphic traits seen in wild populations. Moreover, signals of selection across the genome (as with genomic *Fst* scans) may precisely pinpoint the location of individual loci within these clusters that regulate different components of the selected traits.

In summary, the results from this chapter identify three intervals associated with different aspects of the flower colour in *Antirrhinum*. Moreover, these intervals co-localize with the divergence peaks in *Fst* between samples from the hybrid zone. It is thus desirable to identify candidate genes in these regions, which may be involved in regulating the pigment. The *ROS1* and *ROS2* genes are obvious candidates in the *ROS* region, but nothing is known about the *EL* interval. This is approached in the next chapter, by using gene expression data from different *ROS-EL* haplotypes.

5 Analysis of gene expression in the *ROS-EL* region

The detailed genetic analysis of *ROS-EL* revealed that different components of the flower colour phenotype can be mapped to separate genetic elements in this region. These regions might be under selection, since there is a significant nucleotide divergence between *A. m. pseudomajus* and *A. m. striatum* specifically co-localizing with the mapped loci, a pattern that is not observed in neighbouring regions linked to these loci. This raises the question of which particular genes regulate flower colour to produce the different phenotypes in the two *A. majus* subspecies. Some candidate genes exist for the *ROS* locus, namely *ROS1* and *ROS2*, but not for the newly mapped *EL* region. To find candidate genes controlling flower colour within the mapped intervals in *ROS-EL*, I explored how gene expression differed between flower buds from different genotypes. I approached this by high-throughput RNA sequencing (RNAseq), as it allows sampling the transcripts in a tissue when no prior information about them is available. I identified a *MYB*-like transcription factor (from the same family as *ROS1*) that has higher expression in samples with a dominant *EL* allele compared to samples homozygous for the recessive allele. This gene, denoted *EL-MYB*, falls within the Eluta interval mapped in the previous chapter. Furthermore, gene expression in recombinants within the *ROS* locus (“rosy” recombinants) revealed that regulatory elements 3’ of *ROS1* likely influence its expression. Finally, analysis of polymorphisms between *A. m. pseudomajus* and *A. m. striatum* revealed fixed coding differences in the flower colour genes. These results not only provide with a candidate gene for *EL*, but also highlight that both regulatory and coding changes might be responsible for the flower colour differences between *A. m. pseudomajus* and *A. m. striatum*.

5.1 Introduction: RNAseq datasets

High-throughput RNA sequencing (RNAseq) uses next generation sequence technology to provide a snapshot of the RNA levels of (ideally) all genes expressed in a sample. In this technique, cDNA is obtained by reverse transcription of the RNA sample and sequenced using next-generation platforms (in this work, *Illumina*). The sequencing procedure and raw sequencing data are similar to that obtained from genomic DNA (detailed in section 3.2). However, the nature of RNAseq data differs from that of genomic DNA data, requiring an explanation of how this affects the analysis steps.

Next generation sequencing technologies produce large amounts of short reads, more or less randomly obtained from the DNA (or cDNA) sample being sequenced. In genomic DNA sequencing, a particular base in the genome can be sequenced multiple times (the sequencing depth, *D*; Figure 3.2). In that case, *D* is expected to be more or less homogeneous across the genome (all bases are equally likely to be sequenced, assuming sequencing is random). However, in RNAseq, only the regions corresponding to gene exons are sequenced. Therefore, when reads are mapped to a reference genome, there is unequal *D*, with some regions having no reads at all ($D = 0$, introns and inter-genic regions) and other regions having several reads ($D > 0$, exons) (Figure 5.1). Moreover, some reads will originate from exon-exon boundaries, that is, when mapped to a genome reference, they will be split across introns (this requires the use of a mapping software that supports splitting of reads; methods 2.7.2.2). The high *D* in the exons, together with the split reads across introns, allows the structure of transcripts to be identified without any a-priori information. Therefore, a de-novo assembly of transcripts can be produced from this type of dataset. It is important to perform an assembly step before comparing gene expression between different samples, since such comparison can only be made if a consensus gene set is common between them.

Besides providing a qualitative overview of the transcripts present in a sample, RNAseq also provides quantitative information about the expression level of those transcripts. This relies on the fact that *D* in a particular transcript will be proportional to its abundance in the sample. In other words, a highly expressed gene is more likely to be sequenced than a lowly expressed one, resulting in a difference in *D* between the transcripts of those genes. The consensus way of quantifying expression data from RNAseq is to count how many reads align to a particular transcript. However, a raw measure of read count per transcript is a biased estimate of expression, requiring a normalization step.

The normalization of read counts per transcript is necessary to account for variable gene length and variable size between sequencing datasets. Let us start by considering two genes, A and B, within a sample (Figure 5.1). The read counts of genes A and B are expected to be proportional to their abundance. For example, if gene A has 16 reads and gene B has 8 reads, then one might expect that gene A is expressed twice as much as gene B. However, if gene A is twice the length of gene B, it will have twice as many reads, not because it is expressed more, but simply because it is longer. Therefore, a first step in normalizing read count data is accounting for the length of the transcript, that is, $\frac{read\ count}{length\ transcript}$.

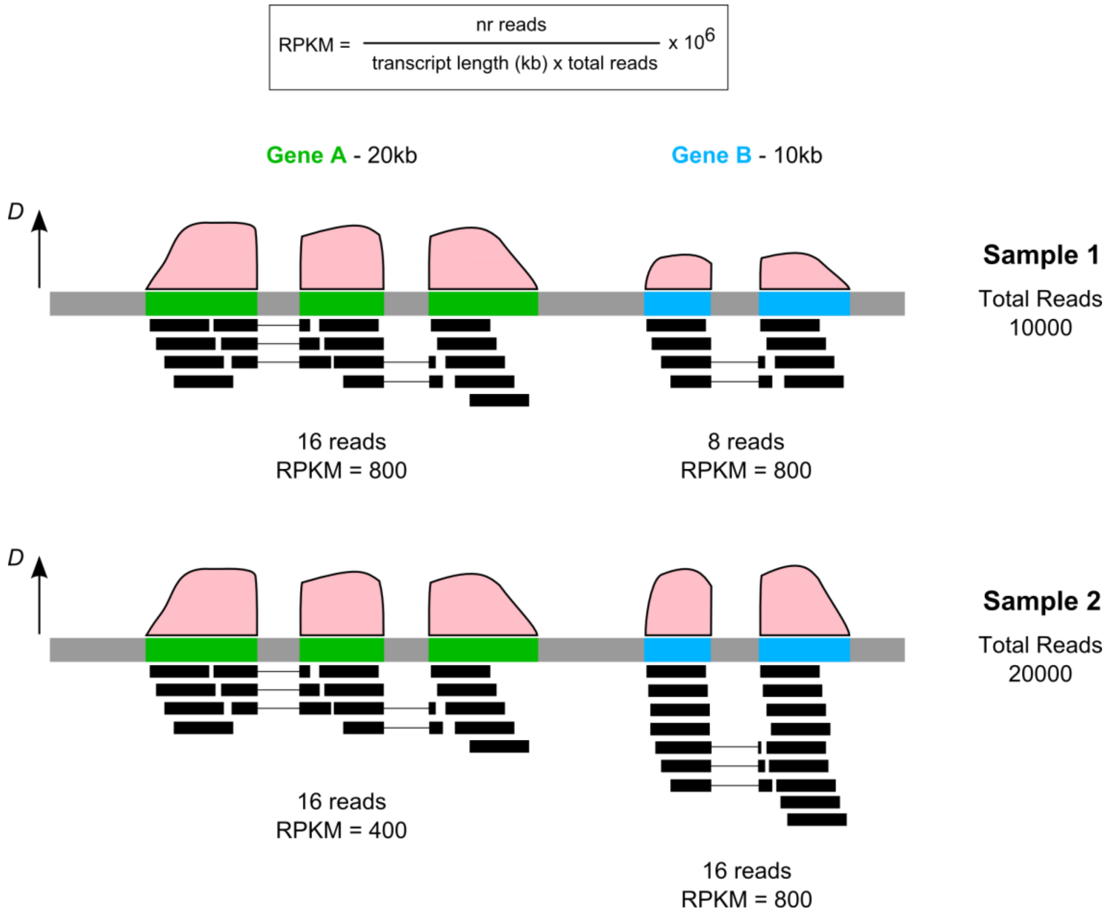


Figure 5.1 – Properties of RNAseq data and an illustration of transcript normalization. Two sequenced samples (1 and 2) are depicted. The grey bar represents the reference genome and the green and blue boxes represent the exons of two genes (A and B). The black boxes are the sequencing reads mapped to the reference genome. Only the exons are sequenced, resulting in abrupt changes in the sequencing depth, *D* (pink histogram). Some reads cross exon-exon boundaries and are therefore split across intron junctions (lines connecting reads). The number of reads mapped to each transcript is expected to be proportional to its expression. However, a normalization procedure is desirable to account for differences in gene size and total number of reads in each sample. One method of normalization is known as *RPKM*, which stands for "Reads Per Kilobase per Million reads" (formula given in figure and text). For example, in sample 1, gene A has twice the number of reads as gene B, but it is also twice longer, therefore both genes can be considered to be expressed at the same level. On the other hand, gene A has the same number of reads in both samples, but sample 2 was sequenced twice as much as sample 1; therefore gene A has half the expression level (RPKM) in sample 2 compared to sample 1.

Another normalizing procedure is required when comparing the expression of genes between samples. For example, if one gene has 8 reads in sample 1 and 16 reads in sample 2 (gene B in Figure 5.1), then we may expect that its expression is doubled in sample 2 in relation to sample 1. However, the two samples might not have been sequenced equally, that is, the total amount of reads may differ between samples. If, for example, sample 1 has a total of 10000 reads and sample 2 a total of 20000 reads, then genes not differentially expressed between samples will have twice as many reads in sample 2 compared to sample 1. Therefore, normalizing for the total read number in each sample is required, that is, $\frac{read\ count}{total\ reads}$.

The two normalizing steps for gene expression, accounting for variable gene length and variable number of total reads in each sample, can be put together in a normalizing equation known as *RPKM*, which stands for "Reads Per Kilobase per Million" (Mortazavi *et al.* 2008), and can be written as:

$$RPKM = \frac{C}{L \times T} \times 10^6$$

Where C is the read count for the transcript, L is the length of that transcript (in kb) and T is the total number of reads in the sample. *RPKM* can therefore be used to compare expression between genes and between samples.

In this chapter I will use *RPKM* as a normalization method for RNAseq data. However, there are alternative normalization methods, as *RPKM* may give biased estimates under certain conditions (Dillies *et al.* 2013). In particular, this method is sensitive if several genes have a very high read-count in some samples and not others. Therefore, *RPKM* is not suitable for statistical analysis of differential expression in such cases. However, I have chosen to use this normalization method due to its intuitive interpretation and to the fact that I do not perform any statistical tests of differential expression. The analyses presented here are mainly exploratory, as I do not have – at this point – replicates which are necessary for the statistical tests of these kinds of data. It should be noted that, even though *RPKM* can be affected by certain aspects of the samples, assuming that most genes across the genome are not differentially expressed between samples of similar genotypes, the biases should not be extreme (Zheng *et al.* 2011; Dillies *et al.* 2013).

The *RPKM* values will be log-transformed (due to the high skewedness of *RPKM* distributions) and will thus be plotted as $\log_{10}$(*RPKM*). The $\log_{10}$ transformation has an intuitive interpretation: a $\log_{10}$(*RPKM*) difference of *n* between two genes means that one gene is 10^n-fold higher than the other gene. Only genes with a minimum read count of 10 in at least one of the samples will be considered for *RPKM* calculation, since there is greater variance on estimated expression for genes with low counts than those with high counts (Zheng *et al.* 2011; Dillies *et al.* 2013).

5.2 Origin of samples and nomenclature

I analysed 11 RNAseq samples from individuals of different *ROS-El* genotypes. It is thus important to clarify the origin of these individuals and establish the nomenclature used throughout the chapter.

I will maintain the nomenclature of the previous chapter to refer to *ROS-EL* haplotypes. Generally, uppercase letters denote a dominant (or semi-dominant) allele and lowercase letters a recessive allele (for example, *ROS el*). When relevant, a superscript is used to denote the origin of each allele (for example a *A. m. pseudomajus* haplotype is *ROS*^{p} *el*^{p}). This nomenclature refers to the subspecies origin of the haplotype itself, whether it is in the original wild genetic background or crossed to another background. Because this chapter deals with the expression of genes (rather than genetically defined loci, which may include several genes), it is important to define the notation for those cases. I will use a similar notation as for genetic loci, for example: *ROS1*^{p} is a dominant *pseudomajus* allele and *ros1*^{s} is a recessive *striatum* allele of the gene *ROS1*. Because of the genetic interactions between *ROS* and *EL,* I will often specify the *ROS-EL* haplotype that the particular allele of the gene is in. For example, *ROS1*^{p} from a *ROS*^{p} *EL*^{s} haplotype and *ROS1*^{p} from a *ROS*^{p} *el*^{p} haplotype refer to the same allele of *ROS1* (*pseudomajus*), but it is in different *ROS-EL* haplotypes. Because *JI7* and *pseudomajus* are identical with regards to their *ROS-EL* haplotypes, I sometimes use the superscript in the form of "7/p" to denote cases where both alleles behave similarly (for example "*ROS1*$^{7/p}$" instead of "*ROS1*^{7} or *ROS1*^{p}").

None of the samples analysed in this chapter are from individuals directly grown from wild capsules. Instead, they are derived from crosses of *A. m. striatum* and *A. m. pseudomajus* to *JI7* and *rosea*^{dorsea} lines (Figure 5.2). These crosses were performed over three generations to establish lines used for the mapping experiments described in chapter 4. Using these individuals has the advantage of allowing comparisons between samples to be made in a similar genetic background, avoiding potential effects of other loci in the genome that may affect the expression of genes in the *ROS-EL* region. For example, if one compares two samples harvested from wild species of *A. majus* and sees a difference in the expression of *ROS1*, it is impossible to tell if that is due to a difference in the *ROS* locus by itself or due to an unlinked trans-acting locus in the genome. The use of individuals sharing a common pedigree should therefore make gene expression comparisons less influenced by modifiers unlinked to *ROS-EL*.

There are two exceptional samples used in this work that do not share the genealogical history described. One is a homozygous *ROS*^{7} *el*^{7} / *ROS*^{7} *el*^{7} individual, which is the actual *JI7* inbred line (not crossed to anything else). The other is a heterozygous *ros*^{s} *EL*^{s} / *ROS*^{7} *el*^{7}, which results from a third-generation backcross of an *A. m. striatum ROS-EL* haplotype to the *JI7* background. However, all samples used have the commonality that most of their genome should be composed of *JI7*, since they have all been crossed to this line several times.

One of the samples was removed from the analysis due to mislabelling. Sample A182-3 (Figure 5.2), thought to be homozygous for an *A. m. striatum* haplotype (*ros1*^{s} *EL-myb*^{s} / *ros1*^{s} *EL-myb*^{s}), was found to carry SNP alleles from a *rosea*^{dorsea} haplotype (like sample A181-32 in Figure 5.2). This sample was thus excluded from the analyses and is not considered in the following sections. In summary, I present expression data for three types of haplotypes: *ROS el*, *ros el* and *ROS EL*.

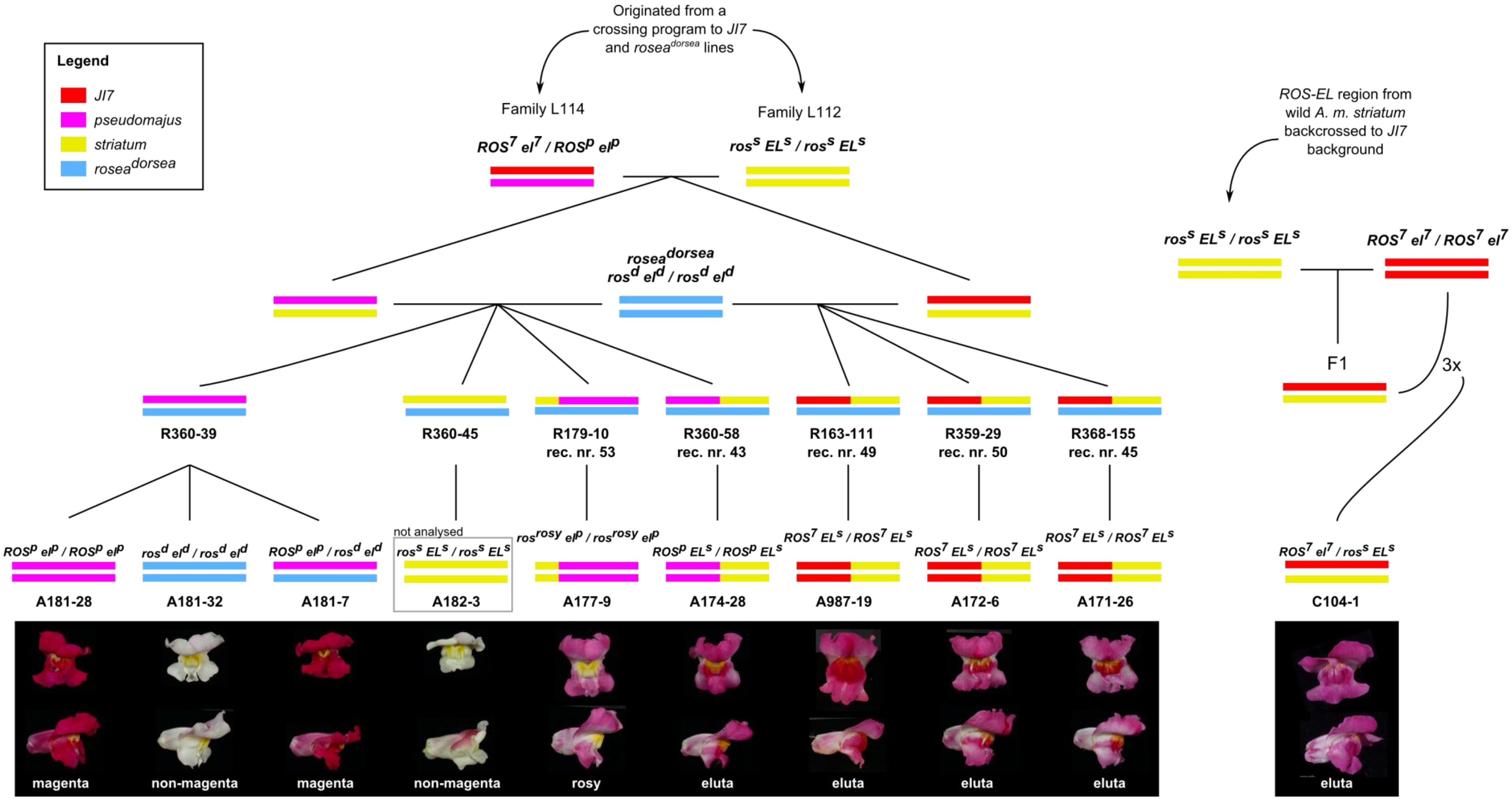

Legend
JI7
pseudomajus
striatum
rosea^{dorsea}
Originated from a crossing program to JI7 and rosea^{dorsea} lines
Family L114
ROS^7 el^7 / ROS^p el^p
Family L112
ros^s EL^s / ros^s EL^s
rosea^{dorsea}
ros^d el^d / ros^d el^d
ROS-EL region from wild A. m. striatum backcrossed to JI7 background
ros^s EL^s / ros^s EL^s
ROS^7 el^7 / ROS^7 el^7
F1
3x
R360-39
R360-45
R179-10
rec. nr. 53
R360-58
rec. nr. 43
R163-111
rec. nr. 49
R359-29
rec. nr. 50
R368-155
rec. nr. 45
ROS^p el^p / ROS^p el^p
ros^d el^d / ros^d el^d
ROS^p el^p / ros^d el^d
not analysed
ros^s EL^s / ros^s EL^s
ros^{rosy} el^p / ros^{rosy} el^p
ROS^p EL^s / ROS^p EL^s
ROS^7 EL^s / ROS^7 EL^s
ROS^7 EL^s / ROS^7 EL^s
ROS^7 EL^s / ROS^7 EL^s
ROS^7 el^7 / ros^s EL^s
A181-28
A181-32
A181-7
A182-3
A177-9
A174-28
A987-19
A172-6
A171-26
C104-1
magenta
non-magenta
magenta
non-magenta
rosy
eluta
eluta
eluta
eluta
eluta

Figure 5.2 – Pedigree of individuals used for RNAseq analysis.

The samples were from the individuals in the bottom-most families (prefix “A” and “C”). The recombinants used for mapping (chapter 4) were from the generation with prefix “R”; the recombinant number is indicated, corresponding to Figure 4.15. Photos of open flowers from each of the individuals used for RNAseq are shown, with the respective phenotype indicated below. The genotypes for *ROS* and *EL* are indicated for each individual. Coloured lines depict the two haplotypes of *ROS-EL*, as indicated in the legend. Sample A182-3 (boxed) was removed from the analysis due to mislabelling (see text).

5.3 Results

5.3.1 Criteria for identifying candidate pigmentation genes from RNAseq

Ignoring the mislabelled sample, there were 10 RNA samples from individuals with different *ROS-EL* genotypes. Each sample was obtained from the corolla of a single flower bud 5-10mm long (Figure 2.1). At this stage, the magenta pigmentation of the flowers is already visible but *ROS1* has high expression, which continues until the flowers open (Schwinn 1999; Schwinn *et al.* 2006). Therefore, 5-10mm seemed a suitable stage to screen for other regulatory genes involved in controlling the pigment of the flowers. The sequencing reads were mapped to the *A. majus JI7* reference genome using a mapping software that allows the splitting of reads across introns (methods 2.7.2.2), resulting in a total of mapped reads that varied between 9.96-97.72 million (Table 5.1). The normalized *RPKM* expression presented here was computed using the *cuffdiff* software (Roberts *et al.* 2011).

Table 5.1 – Summary of RNAseq samples.

Sample	ROS-EL genotype	Total no. of reads (x 10^6)	No. of expressed genes (*RPKM* > 0)
A181-28	$\underline{ROS^{p}\ el^{p}}$ / $\underline{ROS^{p}\ el^{p}}$	26.56	23429
A181-32	$\underline{ros^{d}\ el^{d}}$ / $\underline{ros^{d}\ el^{d}}$	9.96	23278
A181-7	$\underline{ROS^{p}\ el^{p}}$ / $\underline{ros^{d}\ el^{d}}$	22.49	23316
A177-9	$\underline{ros^{rosy}\ el^{p}}$ / $\underline{ros^{rosy}\ el^{p}}$	25.64	22311
A174-28	$\underline{ROS^{p}\ EL^{s}}$ / $\underline{ROS^{p}\ EL^{s}}$	20.26	23686
A987-19	$\underline{ROS^{7}\ EL^{s}}$ / $\underline{ROS^{7}\ EL^{s}}$	22.78	23406
A172-6	$\underline{ROS^{7}\ EL^{s}}$ / $\underline{ROS^{7}\ EL^{s}}$	39.55	24138
A171-26	$\underline{ROS^{7}\ EL^{s}}$ / $\underline{ROS^{7}\ EL^{s}}$	34.31	23254
C104-1	$\underline{ROS^{7}\ el^{7}}$ / $\underline{ros^{s}\ EL^{s}}$	97.72	23356
B7-7	$\underline{ROS^{7}\ el^{7}}$ / $\underline{ROS^{7}\ el^{7}}$	40.55	22874

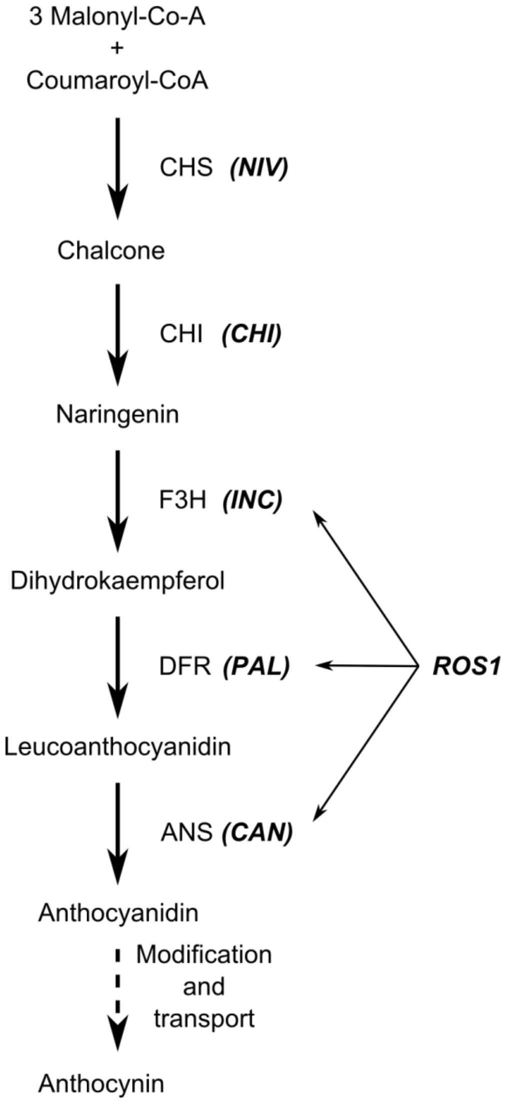


Figure 5.3 – Schematic of the anthocyanin biosynthetic pathway.
The enzymes catalyzing the reactions in each step of the pathway are indicated. The gene names (from *Antirrhinum*) encoding each enzyme are given between parenthesis. Putative regulatory targets of *ROS1* are indicated by arrows pointing away from this gene (according to Schwinn *et al.* 2006). Enzyme abbreviations: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavonol 3-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase. Gene abbreviations: *NIV*, *NIVEA*; *INC*, *INCOLORATA; PAL, PALLIDA; CAN, CANDICA; ROS1, ROSEA1*. Pathway adapted from Martin et al. (1991).

The mapped reads were used to generate a transcript assembly (using the *cufflinks* software; Roberts et al. 2011), resulting in a set of 35267 assembled genes across all samples. Different genes can be expressed in different samples, and therefore the total number of expressed genes in each sample does not reach this number

(Table 5.1). 93 of the assembled genes were located in the *ROS* scaffold, which included *ROS1* and *ROS2*. *ROS3* was not assembled due to its lack of expression in any of the samples. Other known pigmentation genes were also assembled, namely those encoding for enzymes involved in anthocyanin biosynthesis (Martin *et al.* 1985, 1991; Sommer & Saedler 1986): chalcone synthase (*NIV*), chalcone isomerase (*CHI*), flavanone 3-hydroxylase (*INC*), dihydroflavonol 4-reductase (*PAL*) and anthocyanidin synthase (*CAN*) (Figure 5.3). This suggests that the samples are representative of a stage where anthocyanin-related gene expression is taking place.

The distribution of *RPKM* for all expressed genes was similar across all 10 samples, suggesting that the normalization method worked reliably, with no sample having overall different *RPKM* in relation to the others (Figure 5.4). The mean expression level across all samples was *RPKM* = 9.5 and the mode was *RPKM* = 14.0. The number of genes expressed in each sample was similar (~23 thousand; Table 5.1). Gene-by-gene, the correlation of *RPKM* between every pair of samples was high (Pearson's correlation: $0.88 < r < 0.96$), suggesting most genes are not differentially expressed between samples.

Due to the lack of biological replicates, I did not perform any statistical tests for differential expression. Instead, I screened candidate genes in the genetically mapped intervals by considering cases where a gene is reasonably expressed in certain samples (*RPKM* close to or above the average for all genes) but negligibly expressed in others (*RPKM* < 1). This should represent an extreme and significant change of expression, even when no replicates are used. Subtle differences of expression will not be identified with this approach, but, if the genes of interest have major expression differences between genotypes, then they should be successfully identified.

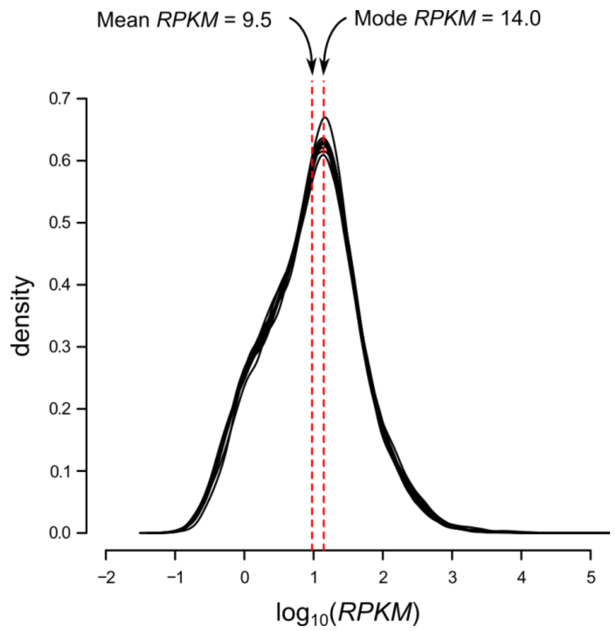


Figure 5.4 – *RPKM* distribution for all genes in the 11 samples analysed.
The distributions are very similar between samples (each sample's distribution is one line), with similar mean and mode (indicated by vertical dashed lines in red). *RPKM* values were log-transformed to attenuate the effect of extremely high and low *RPKM* values and therefore "normalize" the distribution.

To test if this approach was reliable, I first focused on a-priori expectations about *ROS1*. This gene is expressed in the magenta *JI7*, which has a dominant *ROS1* allele, but lowly expressed in the non-magenta *rosea*^{dorsea}, which has a recessive *ros1* allele (Schwinn 1999; Schwinn *et al.* 2006). Therefore, I considered how the expression of *ROS1* changes between samples with different genotypes for this locus, but all carrying a recessive *el* allele: *JI7* and *A. m. pseudomajus* (*ROS1/ROS1*) and *rosea*^{dorsea} (*ros1/ros1*) (Figure 5.5). Both *JI7* and *pseudomajus* samples have a similar expression level of *ROS1*, above the genome-wide mode (*RPKM* = 38 and 49, respectively). Conversely, *rosea*^{dorsea} has comparatively very low levels of *ROS1* expression (*RPKM* = 1.13). This represents at least a 30-fold difference of *ROS1* *RPKM* between these samples (Figure 5.5).

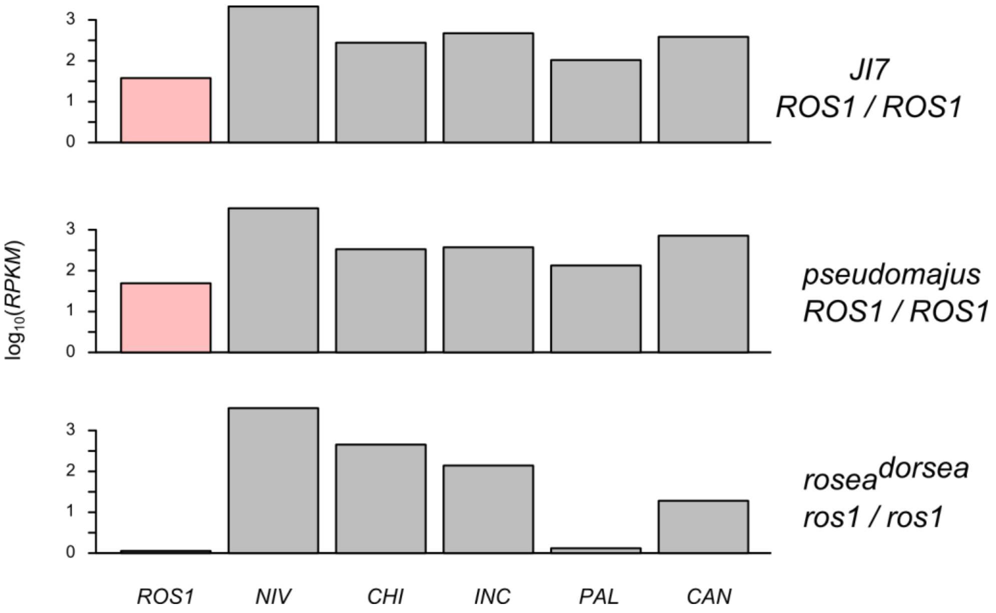


Figure 5.5 – Expression levels of *ROS1* and genes encoding enzymes involved in anthocyanin biosynthesis.
Each barplot panel refers to a particular sample, whose genotype is indicated on the right. The various genes analysed are indicated on the bottom panel and correspond to Figure 5.3.

To find further support for this approach of screening candidate genes, I also looked at the expression of enzymatic genes involved in the anthocyanin pathway (Figure 5.3). *ROS1* regulates genes in the later steps of the pathway – *INC, PAL* and *CAN* – but not those from earlier steps – *NIV* and *CHI* (Schwinn *et al.* 2006). Therefore, the expression of *INC*, *PAL* and *CAN* should be modified like that of *ROS1*, but the expression of *NIV* and *CHI* should be similar between all samples (these can serve as a "control" of between-sample variations).

The enzymatic genes from the earlier steps of the anthocyanin pathway have high *RPKM* in all three samples (NIV and CHI in Figure 5.5). Although there is some variation in *RPKM* values (2130 < *RPKM* < 3529), the differences observed do not correlate with having pigment in the flowers. For example, *rosea*^{dorsea} (*ros1/ros1*) has higher *RPKM* than either *JI7* or *pseudomajus* (*ROS1*/*ROS1*) for both *NIV* and *CHI* genes. This fits with the expectation that *ROS1* expression does not affect the expression of these genes.

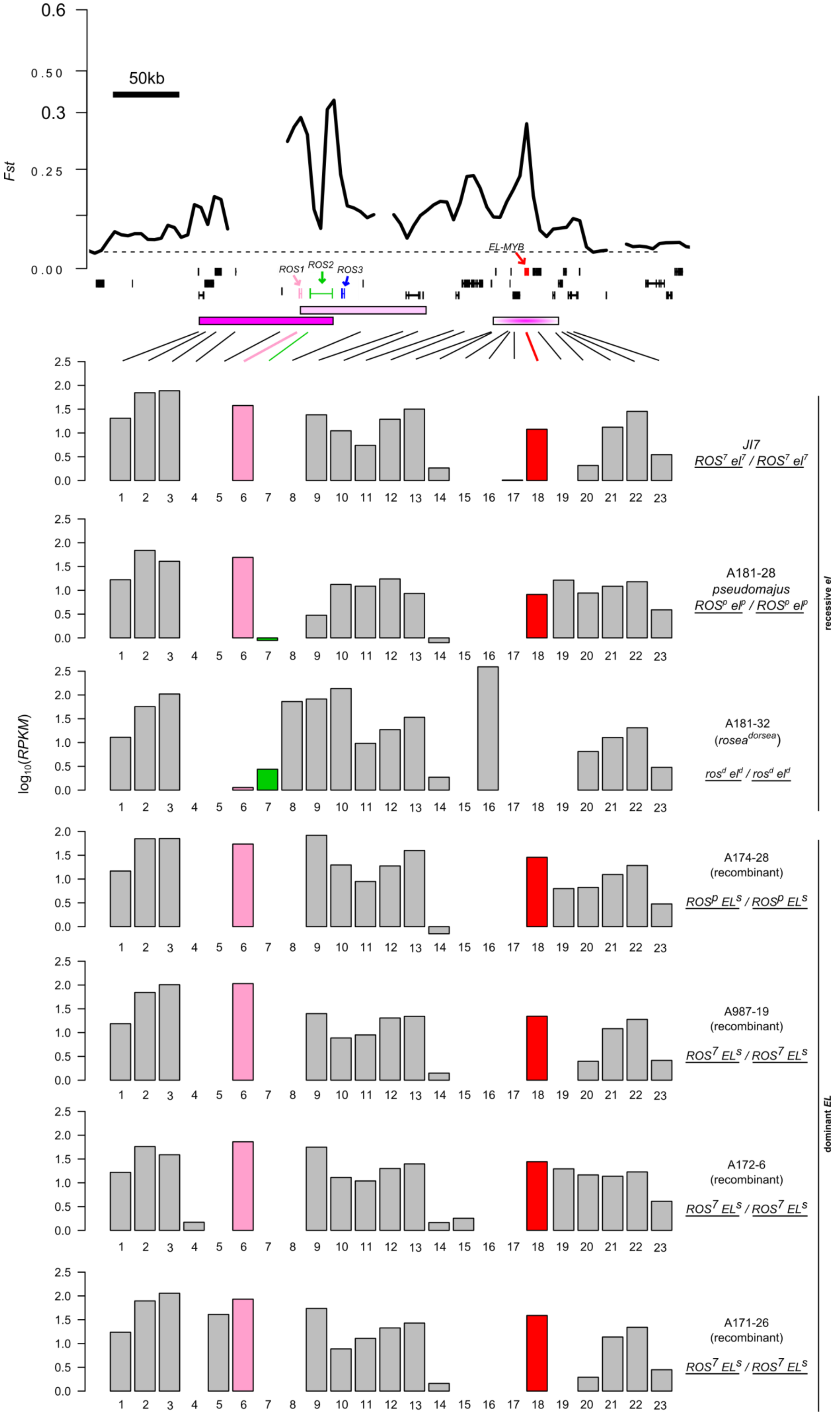
Conversely, the enzymatic genes in downstream steps of the pathway behave more like *ROS1*, having higher expression values in *ROS1/ROS1* relative to *ros1/ros1* genotypes (*INC, PAL* and *CAN* genes in Figure 5.5). *INC* shows the least pronounced change, with only a 2.6-fold difference in *RPKM* between *rosea*^{dorsea} and *pseudomajus* samples. *PAL* and *CAN* have more extreme differences, with higher *RPKM* in *ROS1/ROS1* genotypes (*PAL RPKM* > 100; *CAN RPKM* > 385) than in the *rosea*^{dorsea} *ros1/ros1* genotype (*PAL* RPKM = 1; *CAN RPKM* = 19).

Overall, these results provide good support for the approach of detecting candidate pigmentation genes with major *RPKM* differences between samples. Importantly, *ROS1* showed large (~30 fold) expression differences between samples carrying recessive and dominant alleles in this gene. This test-case suggests that the approach can be reasonably extended to find other candidate genes in the *ROS-EL* region. In particular, it can be extended to find candidate genes associated with different *EL* genotypes (*<u>ROS EL</u>* recombinants compared to *JI7* and *A. m.*

pseudomajus). The caveat of this approach is that subtle (but biologically significant) differences of expression will not be captured, as was the case in the *INC* gene, which would have been discarded as not being different between genotypes.

5.3.2 A candidate gene for EL

In the region defined by the mapped intervals relating to the flower phenotype, there were 23 genes assembled from the 10 RNAseq samples. Using the approach described in the previous section, I looked for candidate *EL* genes by comparing the gene expression in samples with recessive alleles of *el* (*JI7, pseudomajus* and *rosea*^{dorsea}) with that of samples with dominant alleles (four homozygous recombinants $\underline{ROS\ EL}$) (Figure 5.6). These recombinants are not true biological replicates, as each of them originates from a different recombination event. However, the phenotypic consequence of the recombination in all four is similar – all have Eluta flowers –, implying that similar patterns of *ROS1* and *EL-MYB* expression may be expected in all four of them. These four individuals carry a dominant allele of *ROS* – three of them with a *JI7* allele and one of them with a *pseudomajus* allele – together with a dominant allele of *EL* from *striatum* (I refer to them as $\underline{ROS^{7/p}\ EL^{s}}$).



Fst
0.6
0.50
0.3
0.25
0.00
50kb
EL-MYB
ROS1
ROS2
ROS3
log10(RPKM)
2.5
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
JI7
ROS7 el7 / ROS7 el7
A181-28
pseudomajus
ROSp elp / ROSp elp
A181-32
(roseadorsea)
rosd eld / rosd eld
recessive el
A174-28
(recombinant)
ROSp ELs / ROSp ELs
A987-19
(recombinant)
ROS7 ELs / ROS7 ELs
A172-6
(recombinant)
ROS7 ELs / ROS7 ELs
A171-26
(recombinant)
ROS7 ELs / ROS7 ELs
dominant EL

Figure 5.6 (previous page) – Expression level of genes in the *ROS-EL* region in samples with dominant and recessive *EL* alleles.

The top panel shows the window-averaged *Fst*, reprinted from chapter 3. The location of genes is indicated along the x-axis (vertical lines are exons and horizontal lines introns) as well as the mapped intervals corresponding to colour loci represented as boxes (see Figure 4.18). The bottom panel shows barplots of $\log_{10}$(*RPKM*) for each gene located between the mapped intervals for several samples with different genotypes. Some genes are coloured because they are referred to in the text. The individuals' number on the right of each barplot corresponds to Figure 5.2 and their *ROS-EL* genotype is indicated.

As mentioned previously, *ROS1* showed a large *RPKM* difference when comparing *JI7* and *pseudomajus* samples with the *rosea*^{dorsea} sample (transcript 6 in Figure 5.6). All recombinant samples with <u>ROS EL</u> haplotypes expressed *ROS1* at similar levels to *JI7* and *pseudomajus* samples (54 < *RPKM* < 107), as expected from their similar dominant allele of *ROS1*. *ROS2* showed low expression levels in the *rosea*^{dorsea} sample (*RPKM* = 3), and was also detected, at very low levels, in the *pseudomajus* sample (*RPKM* < 1) (transcript 7 in Figure 5.6). This suggests that the pigmentation in *rosea*^{dorsea} flowers might be due to some activity of this gene.

Several genes showed high expression in some samples while being absent in others, but none of these were clearly grouped according to the *ELUTA* genotype (e.g. transcripts 8, 16, 19 in Figure 5.6). Within the mapped Eluta interval, there was one gene with a large expression difference between *EL/EL* and *el/el* samples (transcript 18 in Figure 5.6). This gene's expression values were near average in *JI7* (*RPKM* = 12) and *pseudomajus* (*RPKM* = 8), absent in *rosea*^{dorsea}, but expressed at higher levels in all recombinant samples (22 < *RPKM* < 39). Structuraly, the gene consists of three exons and the predicted protein contains two *MYB* domains, being from the family of *R2R3-MYB* transcription factors, the same family as *ROS1*. A search for homologous proteins (using EBI's *blastp*) retrieved several *MYB*-like proteins, some of them involved in regulating anthocyanin pigment [for example, *TT2* in Arabidopsis (Nesi *et al.* 2001) and *GMYB11* in Gerbera (Laitinen *et al.* 2008)]. No other genes in the region fulfilled such *RPKM* differences for being suitable candidate *EL* genes. This *MYB*-like gene found within the Eluta interval will be referred to as *EL-MYB*.

I took advantage of SNPs in the different alleles of *ROS1* and *EL-MYB* to distinguish which alleles are expressed in a heterozygote with *A. m. pseudomajus* and *A. m. striatum* haplotypes (individual C104-1 in Figure 5.2). In this sample, only the *A. m. pseudomajus* *ROS1*^{p} allele is detected, confirming that *A. m. striatum* has a weak *ros1*^{s} allele. Conversely, both alleles of *EL-MYB* are expressed, which is expected from the results presented for homozygous genotypes. However, it might be expected that the two alleles are not expressed at the same level, since in homozygotes the expression of *EL-MYB*^{s} was ~3 times higher than that of *el-myb*^{p} (Figure 5.6). The presence of SNPs between the two alleles might allow future analysis of allele-specific expression to be made and confirm if this is true.

In summary, the data presented here suggest that the magenta phenotype of *A. m. pseudomajus* results from a high expression of *ROS1* (to promote pigment production) and low expression of *EL-MYB*. Conversely, the non-magenta phenotype of *A. m. striatum* seems to be due to a low expression of *ROS1* and high expression of *EL-MYB*, which presumably inhibits pigment production by any remaining *ROS1* activity (Figure 5.7).

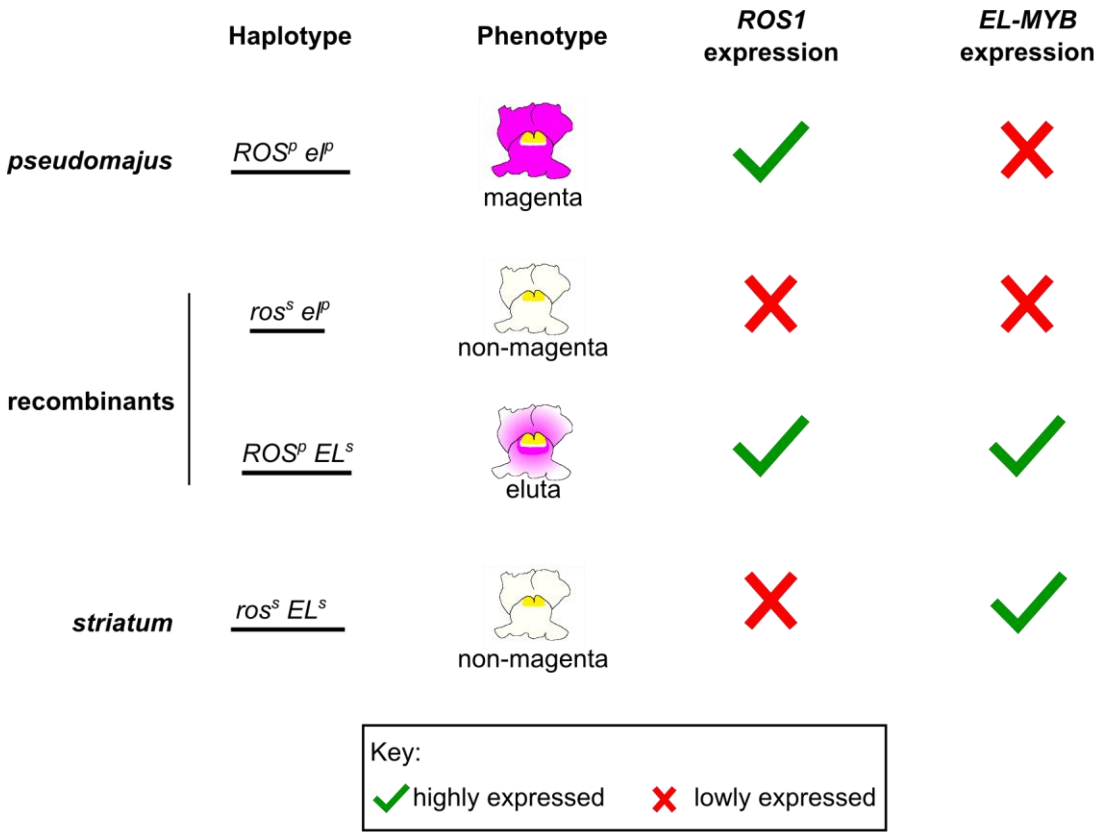


Figure 5.7 – Hypothesis for the expression of *ROS1* and *EL-MYB* genes in *A. m. pseudomajus*, *A. m. striatum* and recombinant *ROS-EL* haplotypes.
Based on the results from this work, high expression of *ROS1* leads to the production of anthocyanin pigment, whereas high expression of *EL-MYB* leads to its repression in particular regions of the flowers.

5.3.3 RNAseq of an individual with rosy phenotype

To find if the paler magenta phenotype of the rosy recombinants was associated with changes in gene expression, I compared the RNAseq data from one $ros^{rosy}\ el^{p}$ / $ros^{rosy}\ el^{p}$ recombinant (individual A177-9; Figure 5.2) with non-recombinant individuals. The rosy haplotype has a recombination point between a marker in the third exon of *ROS1* and the third exon of *ROS2*. It carries a recessive $ros1^{s}$ allele from *striatum* and a recessive el^{p} from *pseudomajus*. Therefore, *ROS1* should have very low expression (similarly to *rosea*^{dorsea} individual) and the same should occur for *EL-MYB* (similarly to *pseudomajus* individual).

The expression results partially fitted this expectation: although the "rosy" recombinant expressed *EL-MYB* at low levels (*RPKM* = 6, comparable with the *pseudomajus* sample with *RPKM* = 8), it expressed *ROS1* at higher levels (*RPKM* = 28, much larger than the *rosea*^{dorsea} sample with *RPKM* ≈ 1) (Figure 5.8). Based on SNPs present in the RNAseq reads, the allele of *ROS1* that is expressed is the one from *striatum*, suggesting that the expression of this allele of *ROS1* changed in the recombinant (as mentioned, in the heterozygote *ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *EL-myb*^{s} the *striatum* copy of *ros1*^{s} was not detected). This could be due to an enhancer of *ROS1* expression located downstream of its coding sequence. As expected from the recombination points, the *EL-MYB* allele being expressed is from *pseudomajus*. The paler magenta phenotype characteristic of the rosy individual was in agreement with a lower expression of both *ROS1* and the enzymatic genes that it regulates: *INC*, *PAL* and *CAN* (Table 5.2).

Table 5.2 – Normalized expression values of *ROS1* and enzymatic genes in the anthocyanin pathway.

The values presented here are the same used in Figure 5.5, except that the rosy sample is now added in. Notice that the expression of *ROS1, INC, PAL* and *CAN* is intermediate in rosy, compared to *rosea*^{dorsea} and *JI7/pseudomajus* samples.

Sample / Gene	*JI7*	*pseudomajus*	*rosy*	*rosea*^{dorsea}
ROS1	38	49	28	1
NIV	2130	3363	4005	3530
CHI	277	335	759	452
INC	472	373	361	139
PAL	104	134	45	1
CAN	386	717	328	19

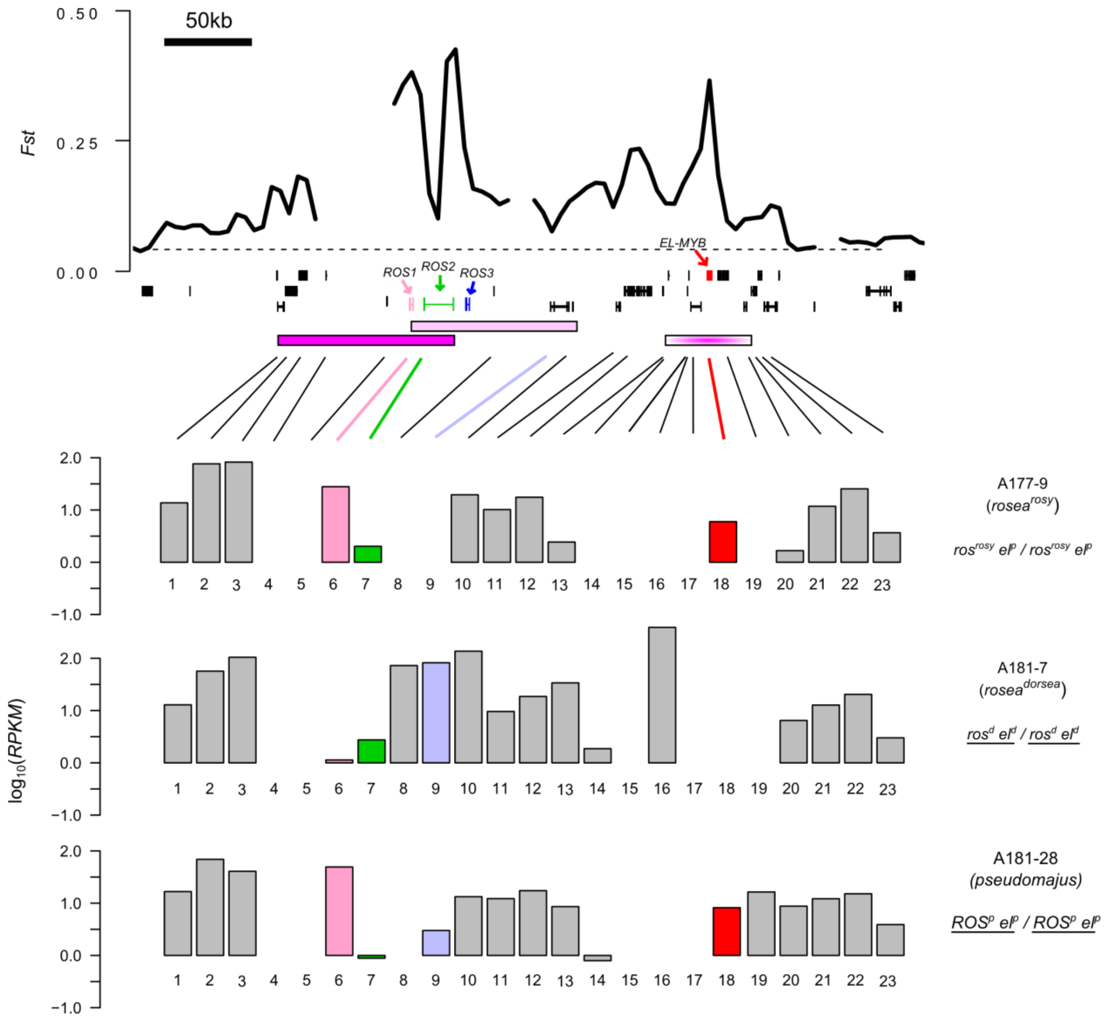


Figure 5.8 – Expression level of genes in *ROS-EL* region in $\underline{ros^{rosy}\ el^{p}}$ and other non-recombinant haplotypes.
Caption as in Figure 5.6.

The $\underline{ros^{rosy}\ el^{p}}$ / $\underline{ros^{rosy}\ el^{p}}$ individual also expresses *ROS2*, although at low levels (*RPKM* = 2). *ROS2* is also expressed at similar levels in *rosea*^{dorsea} (*RPKM* = 2.8), and marginally detected in *pseudomajus* (*RPKM* = 0.8). This suggests that *ROS2* could have a role in the production of low amounts of pigment in these flowers (notice that although *rosea*^{dorsea} is classified as non-mangenta phenotype, it produces some pigment in the upper part of the dorsal petals; Figure 4.1). The molecular markers that defined the recombination in the rosy recombinant used for RNAseq are located in the third exon of *ROS1* and the third exon of *ROS2* (markers 8 and 11 in Table 2.1). Therefore, the exact recombination point should lie between these two markers and cannot be determined without further markers. Because *ROS2* has a large second intron (16.4kb), it is unknown if this individual carries a *pseudomajus*, a *striatum*, or even a recombinant *ROS2* allele. Also, the lack of reads mapping to the *pseudomajus* sample did not allow the determination of the origin of the *ROS2* copy being expressed in this individual based on SNPs.

Finally, there was one gene not expressed in the "rosy" recombinant, but expressed in all other 10 samples analysed (3 < *RPKM* < 83; transcript 9 in Figure 5.8). This gene was located within the mapped interval associated with the "rosy" phenotype. The assembled gene was composed of four exons, and its predicted protein contains a Bet_v_1 domain. This domain is named after the Bet v 1 protein from birch, which is an allergen that causes hay fever in humans (Schenk *et al.* 2009). Despite the conserved domain, there were no highly significant alignments with sequences in public databases. Most alignments were with "major latex-like" proteins, which are proteins with an unknown function, although they have been associated with plant defence and fruit ripening (Ruperti *et al.* 2002; Lytle *et al.* 2009). It is therefore not clear if this gene could be involved in regulating flower colour in *Antirrhinum*.

In summary, the phenotype of this "rosy" recombinant is related to four observations in the RNAseq dataset: moderate expression of a *striatum ROS1* allele;

low expression of a *pseudomajus EL-MYB*; low expression of a *ROS2* allele of undetermined origin; and the lack of expression of a gene of unknown function.

5.3.4 Screening putative functional mutations in colour genes

One hypothesis for the function of *EL-MYB* is that it acts as a repressor of magenta pigment. However, the fact that both *JI7* and *pseudomajus* samples express *EL-MYB* (albeit at lower levels than dominant *EL* alleles) but are nonetheless magenta, raises the question of whether certain mutations alter the function of their proteins. I investigated if this is a plausible explanation by analysing SNPs and indels obtained from the RNAseq sequences of the *EL-MYB* transcript of *striatum* (which is expressed in <u>*ROS EL*</u> */* <u>*ROS EL*</u> samples) and *pseudomajus*, and compared both with the reference sequence from *JI7*.

In the coding sequence of *EL-MYB* (spread over three exons) I found several nucleotide differences between *JI7*, *pseudomajus* and *striatum*, some of which resulted in amino acid differences at the protein level (Figure 5.9A). Several polymorphisms were *striatum*-specific, and mostly located in the C-terminus of the protein, downstream of the MYB domains. One exception was a change from a Methionine (M) to a Threonine (T) in the second MYB domain. There was also a 3bp nucleotide insertion in the coding sequence, which resulted in an extra Lysine (K), located just before the start of the first MYB domain. Finally, a 1bp insertion near the C-terminus resulted in a frame-shift with a premature stop codon. The *JI7* and *pseudomajus* protein sequences were very similar, with two polymorphisms being specific to each of those sequences (Figure 5.9A). While in *JI7*, both polymorphisms were located downstream of the MYB domains, in *pseudomajus* the two amino-acid differences were located in the MYB domains.

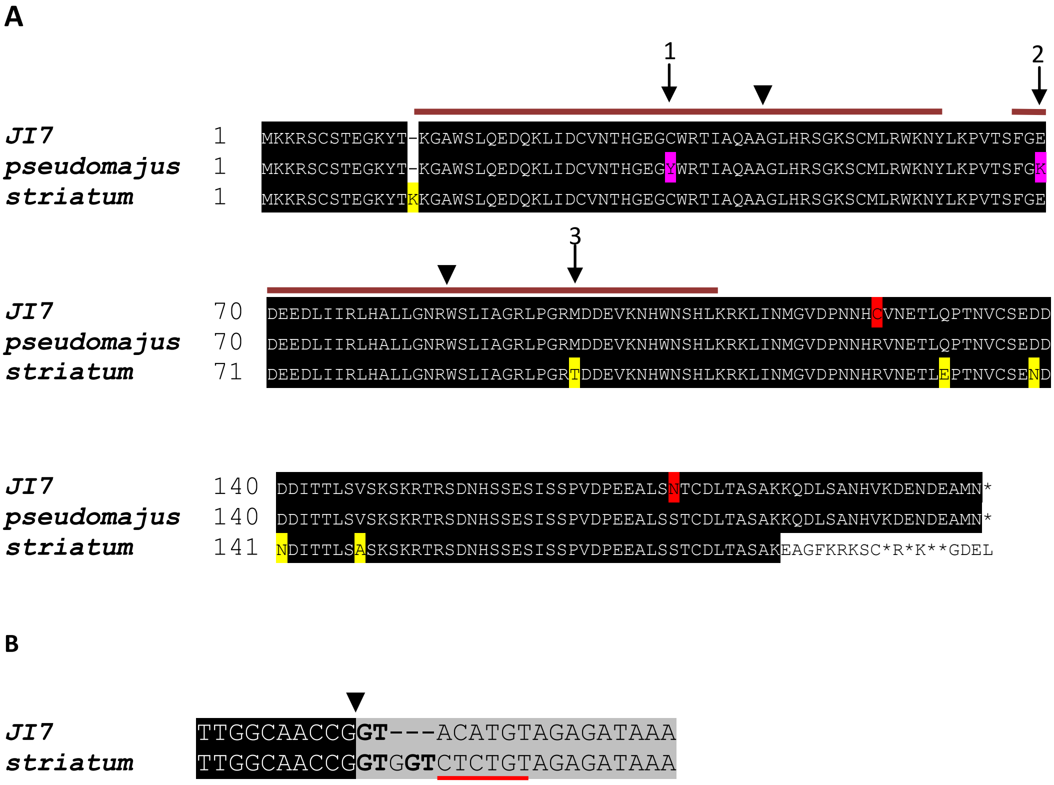

A
1
2
JI7 1 MKKRSCSTEGKYT-KGAWSLQEDQKLIDCVNTHGEGCWRTIAQAAGLHRSGKSCMLRWKNYLKPVTSFGE
pseudomajus 1 MKKRSCSTEGKYT-KGAWSLQEDQKLIDCVNTHGEGYWRTIAQAAGLHRSGKSCMLRWKNYLKPVTSFGK
striatum 1 MKKRSCSTEGKYTKKGAWSLQEDQKLIDCVNTHGEGCWRTIAQAAGLHRSGKSCMLRWKNYLKPVTSFGE
3
JI7 70 DEEDLIIRLHALLGNRWSLIAGRLPGRMDDEVKNHWNSHLKRKLINMGVDPNNHCVNETLQPTNVCSEDD
pseudomajus 70 DEEDLIIRLHALLGNRWSLIAGRLPGRMDDEVKNHWNSHLKRKLINMGVDPNNHRVNETLQPTNVCSEDD
striatum 71 DEEDLIIRLHALLGNRWSLIAGRLPGRTDDEVKNHWNSHLKRKLINMGVDPNNHRVNETLEPTNVCSEND
JI7 140 DDITTLSVSKSKRTRSDNHSSESISSPVDPEEALSNTCDLTASAKKQDLSANHVKDENDEAMN*
pseudomajus 140 DDITTLSVSKSKRTRSDNHSSESISSPVDPEEALSSTCDLTASAKKQDLSANHVKDENDEAMN*
striatum 141 NDITTLSASKSKRTRSDNHSSESISSPVDPEEALSSTCDLTASAKEAGFKRKSC*R*K**GDEL
B
JI7 TTGGCAACCGGT---ACATGTAGAGATAAA
striatum TTGGCAACCGGTGGTCTCTGTAGAGATAAA

Figure 5.9 – Sequence alignments of *EL-MYB* from *JI7*, *pseudomajus* and *striatum*.

A) Protein alignment translated from the coding sequence of *EL-MYB***.** Brown lines above the alignment indicate the position of the two MYB domains. Arrow heads point to the location of introns in the genomic sequence. Numbered arrows point to polymorphisms found in the MYB domains (numbers correspond to Table 5.3). The misalignment in the C-terminus of the protein is due to a frame-shift in the *striatum* allele. Polymorphic amino-acids are coloured as: yellow – *striatum* specific; magenta – *pseudomajus* specific; red – *JI7* specific.

B) Nucleotide alignment showing the 5' splice junction of the second intron of *EL-MYB*. Sequences in black denote the exon and those in grey the intron. Arrow head points to the splice site. Nucleotides underlined in red, point to a GGT duplication in the *striatum* allele. Notice that this duplication originates a new GT sequence (in bold font), which is a conserved sequence for intronic 5' splice sites.

Some RNAseq reads in *ROS EL*/*ROS EL* samples mapped to what were predicted to be the introns of *EL-MYB* in the *A. m. striatum* sample. This could suggest alternative splicing of the *EL-MYB* mRNA, resulting in the expression of different isoforms of this gene. Unfortunately, the number of reads covering the introns was low, precluding any conclusions about its relative expression in different haplotypes. However, based on those few reads, I was able to detect a polymorphism in the splice junction of the second intron of *EL-MYB* (Figure 5.9B). Usually, there is a conserved GT sequence at the 5' end of the introns and an AG sequence at the 3' end (Alberts *et al.* 2002). At the 5' end of the second intron of the *striatum EL-MYB*, the GT sequence is present, but immediately followed by a GGTCTC sequence instead of ACA found in *JI7* (Figure 5.9B). It could be that this change adjacent to the 5' donor splice site of the intron perturbs the mRNA splicing of this gene, resulting in alternative *EL-MYB* transcripts.

To investigate if the observed amino-acid differences in the MYB domains, could affect the function of the protein, I aligned the protein sequences of *JI7*, *pseudomajus* and *striatum* with homologous proteins obtained from a BLAST search. My assumption is that changes that do not affect the function of the protein should be shared with other homologous proteins, whereas changes that disrupt the protein's function should not. I aligned the MYB domains of my three sequences with the MYB domains of 50 homologous proteins (only MYB domains were used because the proteins are highly divergent elsewhere). I focused attention on the three amino-acid polymorphisms that were found between *JI7*, *pseudomajus* and *striatum* (numbered in Figure 5.9A). The first two polymorphisms are *pseudomajus*-specific, that is, *JI7* and *striatum* shared the same amino-acid. One of them (a tyrosine [Y] in the first MYB domain) was not found in any other homologous protein, and the other (a lysine [K] in the second MYB domain) was found in only 3 other proteins (Table 5.3). In both cases, the amino-acids from *JI7* and *striatum* were the most commonly found in the 50 homologous sequences (Table 5.3). The third polymorphism was *striatum*-specific (a threonine [T] in the second MYB domain), but it was absolutely conserved in all 50 homologous sequences, whereas *JI7* and *pseudomajus* (both with a methionine [M]) differed from all others. This suggests that the *striatum* sequence is in fact more similar to other homologous sequences, and that *JI7* and *pseudomajus* differ the most.

Table 5.3 – Comparison of amino-acid composition of *JI7*, *pseudomajus* and *striatum* with 50 homologous proteins for three polymorphisms found in the MYB domains of *EL-MYB*. The three polymorphisms correspond to those highlighted in Figure 5.9A. For each polymorphism, all amino-acids found amongst the compared proteins (50 from BLAST, plus 3 from *A. majus*) are shown, with their respective count. Cells in magenta in polymorphisms 1 and 2 refer to the *pseudomajus* amino-acid and cells in grey to the *JI7* and *striatum* amino-acid. The cell in yellow in polymorphism 3 refers to the amino-acid in *striatum* and the cell in grey to the amino-acid in *JI7* and *pseudomajus*.

Polymorphism 1	
amino-acid	**Count**
C	16
G	7
N	7
K	6
R	5
V	4
S	2
H	2
Q	2
Y	1
L	1

Polymorphism 2	
amino-acid	**Count**
E	14
P	11
S	6
D	5
A	4
Q	3
Y	3
K	3
L	2
N	1
H	1

Polymorphism 3	
amino-acid	**Count**
T	51
M	2

Based on these results, I explored if these putative functional mutations in *EL-MYB* occurred as fixed differences between *A. m. pseudomajus* and *A. m. striatum* populations from the hybrid zone. The pooled sequencing data of hybrid zone samples (Chapter 3) were used to find all nearly-fixed (Δp > 0.8) SNPs between the two furthest sequenced pools (YP4 and MP11; Table 3.1). More than half of these SNPs were located in the vicinity of the colour genes, *ROS1* to *ROS3* and *EL-MYB*, coincident with the *Fst* peaks (Figure 5.10). From those SNPs that were located in the coding sequence (CDS) of these genes, I further characterized them as synonymous (if they did not result in an amino-acid change) or non-synonymous (if they did result in an amino-acid change) (Figure 5.10B). No nearly-fixed SNPs were found in the CDS of *ROS3.* In the other genes, several changes in the CDS were non-synonymous between populations, suggesting some protein divergence between the two subspecies. In *ROS1*, the two nearly-fixed amino-acid changes occurred in the 3'-region of the protein, downstream of the MYB domains; in *ROS2*, there were two non-synonymous changes in each of the MYB domains and three changes further downstream; finally, in *EL-MYB*, there was one mutation in the second (3'-most) *MYB* domain and three mutations further downstream.

Two of the amino-acid differences in the MYB-domains analysed in the RNAseq data (numbered 1 and 2 in Figure 5.9 and Table 5.3) were not diagnostic differences between the hybrid zone pools. However, both of them were fixed in the *A. m. striatum* sample, while being polymorphic in the *A. m. pseudomajus* sample (at a frequency of 0.3 and 0.1, respectively). Conversely, the amino-acid difference specific to *A. m. striatum* (number 3 in in Figure 5.9 and Table 5.3) was fixed between samples (*Δp* > 0.99). Thus, the alleles analysed in the RNAseq data are not shared between *A. m. striatum* and *A. m. pseudomajus*, and thus might have some functional significance for the production of anthocyanin pigment.

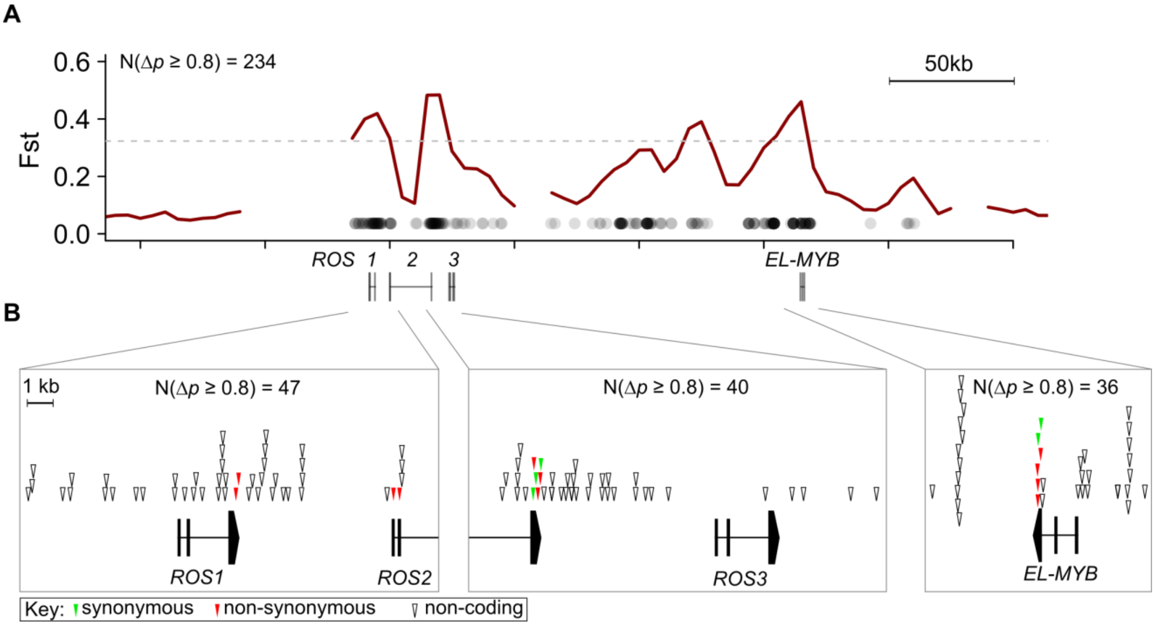


Figure 5.10 – Location of fixed polymorphisms around putative flower colour genes.

A) *Fst* (averaged across 10kb windows) between the two furthest pools sequenced from the hybrid zone (YP4 and MP11). The points along the x-axis denote the location of nearly-fixed polymorphisms between the two samples (defined as SNPs with an allele frequency difference *Δp* ≥ 0.8). Points are coloured with transparency to denote their density. The dashed line is the 99.5th quantile of the whole-genome *Fst* distribution. The position of the colour genes focused on in this work is indicated below the plot. The total number of nearly -fixed polymorphisms in the region is given above the plot.

B) Location of nearly-fixed polymorphisms (triangles) in relation to the coding sequences of *ROS1*, *ROS2*, *ROS3* and *EL-MYB* genes. Coding sequences are shown as boxes and introns as lines. The number of nearly -fixed polymorphisms is given in each box. SNPs are coloured as indicated in the key.

5.4 Discussion

5.4.1 *EL-MYB* is a likely candidate for *EL*

The comparison of gene expression between samples with recessive and dominant alleles of *EL*, revealed a candidate gene for *EL* within the genetically mapped Eluta interval. This gene (*EL-MYB*) was found to encode a protein with an *R2R3-MYB* domain, the same type as *ROS1*, which made it a candidate for being involved in the Eluta phenotype of *Antirrhinum*. *EL-MYB* shares similarity with other genes involved in anthocyanin biosynthesis, namely in fruits (e.g. *MYB4* in *Vitis vinifera*; Matus et al. 2009), seeds (e.g. *AtTT2* in *Arabidopsis thaliana*; Nesi et al. 2001) and flowers (e.g. *GMYB11* in *Gerbera hybrid*; Laitinen et al. 2008).

The magenta phenotype of *Antirrhinum* flowers seems to involve the expression of *ROS1* at relatively high levels (compared to *rosea*^{dosea}) and *EL-MYB* at lower levels (compared to *ROS EL* recombinants) (Figure 5.6). Conversely, in *ROS EL* recombinants both *ROS1* and *EL-MYB* are expressed at high levels, and this correlates with the restricted pigment observed in their flowers. This suggests that the EL-MYB protein might act as a repressor of pigment production promoted by *ROS1*. Indeed, some of the homologs of *EL-MYB* have a repressor activity in the production of phenolic compounds, including anthocyanins (e.g. Jin et al. 2000; Aharoni et al. 2001), giving some support to this hypothesis. The mechanism for this repression remains to be clarified, but it does not seem to involve a direct regulation of *ROS1* expression, otherwise its expression would be lower in *ROS EL* recombinants, which was not observed. Another hypothesis is that the EL-MYB protein competes with the ROS1 protein for DNA targets and/or in the formation of protein complexes necessary for downstream activation of target genes (Ramsay & Glover 2005). Indeed, the maize *C1* gene (*ROS1* homolog) is known to bind to a bHLH transcriptions factor (*B*), and the formation of this complex is necessary for anthocyanin production (Goff *et al.* 1992). Thus, EL-MYB might compete with ROS1 in the formation of regulatory protein complexes, leading to the repression of pigment in particular parts of the flower where it is expressed. This hypothesis could be investigated in the future, for example, by doing yeast-two-hybrid assays to investigate if ROS1 and EL-MYB proteins interact with the same set of proteins or with each other.

The observation that *JI7* and *A. m. pseudomajus* samples express *el-myb* (although at lower levels than in *ROS EL* recombinants) and yet have a magenta phenotype requires an explanation. One hypothesis is that the *striatum* and *pseudomajus* EL-MYB proteins are functionally different. I found several polymorphisms between *JI7*, *pseudomajus* and *striatum*, which could be candidates for altering the function of the EL-MYB protein (Figure 5.9). I focused attention on three polymorphisms in the MYB domains of *EL-MYB*, since these contain the DNA-binding motifs of the protein (Dubos *et al.* 2010). Three amino-acid polymorphisms in this region revealed that *JI7* and *pseudomajus* are more distinct from homologous proteins available in protein databases than *striatum* is (Table 5.3). In particular, a polymorphism in the second MYB domain (polymorphism 3 in Figure 5.9A and Table 5.3) was found to be conserved in all 50 homologous sequences considered, but different in *JI7* and *pseudomajus*. This polymorphism was also found to be fixed between samples in the hybrid zone (Figure 5.10). Together, these results suggest that *JI7* and *pseudomajus* proteins might be non-functional, despite some *EL-MYB* expression in the flowers.

All of the functional hypotheses discussed here are of a speculative nature, since they are based on limited molecular data. However, they point the way for future work on *EL*. For example, the role of *EL-MYB* in regulating pigment production in flowers may be elucidated by its expression pattern in the floral tissue, namely by *in situ* hybridization. If the dominant *EL-MYB* allele represses the pigment in the outer lobes and tube of the flowers, then it might have an expression pattern within those regions of the flower bud. Currently, an *in situ* probe for the *JI7 EL-MYB* allele has been developed and a preliminary in-situ trial tested on *JI7* flower buds, but the results are still inconclusive (João Raimundo, pers. comm.). Futhermore, it would be useful to obtain knock-out alleles of the *striatum EL-MYB* allele in *ROS EL* recombinants, which should result in magenta flowers.

5.4.2 Analysis of rosy phenotype reveals complexity of *ROS* locus

Many of the nearly-fixed polymorphisms in the *ROS* scaffold between *A. m. pseudomajus* and *A. m. striatum* in the hybrid zone were found within the vicinity of the *ROS* genes (Figure 5.10). Some of these polymorphisms cause non-synonymous changes in the predicted proteins of the genes and thus might be associated with functional differences between the encoded proteins. However, most of these polymorphisms occur in non-coding regions of the colour genes (Figure 5.10B), suggesting that the phenotypic difference between the subspecies is also due to cis-regulatory mutations. The analysis of RNAseq data from an individual with the rosy phenotype (*ros*^{rosy} *el*^{p}/*ros*^{rosy} *el*^{p}) revealed a unique profile of gene expression in the *ROS* locus, which provided support to this hypothesis. Although this rosy individual carries a *ros1*^{s} allele (which is not detected in *ROS1*^{p} *el-myb*^{p} / *ros*^{s} *EL*^{s} heterozygotes), this gene is expressed at a level near to that of *JI7* and *pseudomajus* samples (Figure 5.8). Therefore, the recombination in this individual – between the third exon of *ROS1* and the third exon of *ROS2* – modified the expression of the *ros1*^{s} allele. This suggests that regulatory elements downstream of the *ROS1* coding sequence influence the expression of this gene. Cis-regulatory mutations are often argued to be major contributors for phenotypic differences between species, because they have the potential to avoid functional trade-offs due to a gene's pleiotropic activity (Carroll 2005; Wray 2007). For example, if *ROS* genes are involved in regulating anthocyanin production in other tissues besides petals (this is currently unknown), then changes in floral pigmentation can be accomplished by regulatory changes that do not affect the synthesis of anthocyanins in other tissues (such as leaves, as is often observed in both *A. m. striatum* and *A. m. pseudomajus* plants). Dissecting how coding and cis-regulatory mutations contribute to functional and phenotypic divergence is challenging, particularly in non-model organisms where certain molecular tools (such as transgene expression) are not available. However, this work shows how combining information from naturally-occurring polymorphisms, gene expression and recombinant mapping allows tackling this question.

The rosy recombinant also expresses the *ROS2* gene, although with lower *RPKM* than *ROS1* (Figure 5.8). This suggests that *ROS2* may also contribute to the pigment seen in rosy flowers. *ROS2* was also found to be expressed in *rosea*^{dorsea}, which has low amounts of pigment in the outer epidermis of the dorsal petals (Figure 4.1). Although *ROS2* was previously found not to be expressed in *rosea*^{dorsea} (Schwinn *et al.* 2006), the RNAseq data suggest that it may be expressed at low levels and perhaps contribute to the phenotype of these flowers. This suggests that both *ROS1* and *ROS2* genes may be involved in regulating anthocyanin production in flowers of *Antirrhinum*. To understand how these different genes contribute to the final phenotype of the flowers, it would be desirable to see when (in developmental time) and where they are expressed. As with *EL-MYB*, two probes for *ROS1* and *ROS2* have been constructed to perform *in situ* hybridizations in *JI7* flowers, but this is still underway. It would also be interesting to know if the *ROS3* gene is functional or is a non-functional pseudogene.

Finally, the *ROS* locus may provide an interesting case for studying the evolution of gene function in duplicated genes (Zhang 2003). The high sequence similarity between *ROS1*, *ROS2* and *ROS3* and their close location in the genome suggest that these genes are paralogues resulting from gene duplication events. This provides the opportunity to see if they evolved to regulate pigment in different tissues, if their expression differs between species of *Antirrhinum* with distinct flower colours, or even if they occur in closely-related species, such as *Linaria* or *Mimulus*. Extending the analysis to other species might help to determine when this duplication occurred and what the role of the *ROS* homologs is in regulating floral pigment in those species.

5.4.3 Effectiveness of RNAseq to find candidate genes involved in flower colour

RNAseq is a powerful method for accessing the pool of transcripts present in a particular tissue, when little knowledge about those transcripts is available. This is the case in *Antirrhinum*: the genome sequencing project is in its early days, and information about gene location and its functional annotation (i.e. biological knowledge about those genes, such as conserved domains, homology to other genes, etc.) is scarce. Therefore, RNAseq seemed a suitable approach to explore changes in gene expression associated with the mapped genetic components of flower colour (chapter 4), allowing the identification of candidate genes within those mapped regions. The RNAseq dataset used in this study, was obtained from 10 samples (Table 5.1) and allowed the assembly of 35267 genes. This number is slightly lower than the predicted number of genes in the *A. majus* genome (47555 genes; Yongbiao et al., unpublished). However, the assembly obtained from my dataset is likely to be incomplete, since it is based on RNA from a single tissue (corolla of flowers).

I have opted to use RNA samples extracted from corollas of a single flower bud and from a single developmental window (5-10mm long buds). The choice of stage was based on *ROS1* expression, which is known to be significant at this developmental stage (Schwinn *et al.* 2006). However, other genes controlling flower colour may be expressed at different stages of development and therefore this strategy may not capture all of the gene expression variation associated with the phenotypes studied here. Several buds of different sizes could have been pooled for each individual to make the RNA extraction. One potential caveat of this approach is that it may introduce stages of development where the genes associated with flower colour are not expressed at all. This would result in a "dilution effect" of those genes of interest, potentially hindering their detection in the RNAseq data. Although I cannot confidently say which would have been the optimal strategy, the assembly of several known genes involved in anthocyanin biosynthesis (including *ROS1* and *ROS2*), is indicative that the sampled tissues represent a stage in development suitable to screen for other regulatory genes controlling flower colour.

One clear limitation of this dataset is that no biological replicates were sequenced for each of the *ROS-EL* genotypes analysed, which is essential for conducting statistical analysis of differential expression (Dillies *et al.* 2013). Replicates were not included in this work, because it constituted a first attempt to apply this method to screen candidate genes associated with flower colour changes related to the *ROS-EL* genotype. Given the promising results presented here, two extra biological replicates are now being sequenced for some of the samples described, as well as for a sample homozygous for an *A. m. striatum* haplotype. This should consolidate some of the results obtained, allowing a more quantitative description of changes in gene expression.

Due to the lack of replicates, I opted to use an exploratory analysis that only considered extreme changes of normalized gene expression (*RPKM*) between samples of different *ROS-EL* genotypes. This approach is limited, since significant changes in expression associated with flower colour might be dismissed (false negatives) and others might be wrongly considered (false positives). To see if the approach was reasonable, I looked at the expression data from a set of "test" genes

– *ROS1*, *INC, PAL* and *CAN* (Figure 5.3) – whose expression is expected to differ between different *ROS1* genotypes (Schwinn *et al.* 2006). The results largely fitted with the expectation that *ROS1* is expressed in *ROS1/ROS1* individuals but not in *ros1/ros1* (except in a rosy individual, discussed earlier) (Figure 5.5). Furthermore, the genes regulated by *ROS1* showed expression levels that correlated with the expression of this gene (Figure 5.5). This indicates that my exploratory approach could correctly pinpoint genes involved in flower colour, as long as the differences in gene expression are large (for example, one of these "test" genes – *INC* – might have been dismissed as involved in flower colour if not already known).

Although many questions remain unanswered, the results reported in this chapter and the previous one reveal that intricate interactions between loci in the *ROS-EL* region are involved in the regulation of flower colour. This chapter analysed the expression of three genes – *ROS1, ROS2* and *EL-MYB* – that are likely important for the flower colour differences seen between *A. m. pseudomajus* and *A. m. striatum*. The consequences of recombination between the *A. m. pseudomajus* (<u>*ROS el*</u>) and *A. m. striatum* (<u>*ros EL*</u>) haplotypes are clear both at the phenotypic and gene expression levels. These recombination events are relatively easy to observe in controlled crosses made in the glasshouse. Therefore, this raises the question of whether natural recombinants occur in the hybrid zone where the two subspecies haplotypes co-occur. If that is the case, what are the consequences for the phenotype of the flowers? I approach this question in the next chapter, focusing attention on the three genes studied in this chapter (*ROS1*, *ROS2* and *EL-MYB*).

6 Characterization of natural *ROS-EL* recombinants in the *pseudomajus* x *striatum* hybrid zone

Genetic analysis of flower colour in *A. m. pseudomajus* and *A. m. striatum* has shown that the *ROS-EL* region is responsible for the major difference in the floral magenta pigmentation between these subspecies (Whibley 2004; Whibley *et al.* 2006). In particular, *A. m. pseudomajus* is characterized by a *ROS el* haplotype and *A. m. striatum* by a *ros EL* haplotype. The genetic mapping experiments in this work allowed the identification of an interval containing *EL* and the dissection of sub-regions within the *ROS* locus controlling different aspects of floral pigmentation. Additionally, the gene-expression experiments allowed identification of three genes that control the pigmentation differences between *A. m. pseudomajus* and *A. m. striatum*: *ROS1* and *ROS2* at the *ROS* locus and *EL-MYB* at the *EL* locus. At the population level, these loci (and their corresponding genes) are highly divergent (high *Fst*) between samples collected from the flanks of the *pseudomajus* x *striatum* hybrid zone. This suggests that the subspecies allelic combination of the *ROS-EL* region (*ROS el* and *ros EL*) is maintained by selection in the parapatric populations. However, in the central region of the hybrid zone, where the two subspecies haplotypes can occur in a heterozygous form, there are opportunities for recombination to occur between the *ROS el* and *ros EL* haplotypes. I used molecular markers in *ROS1* and *EL-MYB* to genotype samples across the hybrid zone transect, and show that recombinant haplotypes (*ROS EL* and *ros el*) occur at ~5% frequency in the population. Using controlled crosses, these recombinants were confirmed to change the phenotype of the flowers, as expected from the genetic experiments in previous chapters. These crosses also revealed that other modifiers of pigmentation might be segregating in the hybrid zone. Finally, *EL* was shown to interact with another component of petal coloration: vein-specific pigmentation (venation). Dominant alleles of *EL* restrict the venation pattern in the dorsal petals of the flowers, whereas in *el/el* individuals the venation is spread in those petals. These results reveal how particular patterns of colour in *A. m. pseudomajus* and *A. m. striatum* require the interaction between the *ROS* and *EL* loci. This is discussed in view of the different forms of selection that might maintain each subspecies' haplotypes (*ROS el* and *ros EL*), namely co-adaptation (epistatic selection) or frequency-dependent selection.

6.1 Introduction

Using controlled crosses in the glasshouse between *A. m. pseudomajus* and *A. m. striatum*, I was able to show that recombination between *ROS el* and *ros EL* haplotypes is not suppressed in these subspecies. However, the two loci are tightly linked (~0.5cM apart), which required sowing several hundreds of individuals from controlled crosses in order to observe a significant number of recombination events between *ROS* and *EL*. In the context of the *pseudomajus* x *striatum* hybrid zone, one could hypothesise that recombination between *ROS* and *EL* is so rare that it precludes the accumulation of a significant number of recombinant haplotypes in the population. However, this expectation might be false, as the accumulation of recombinants in a population is not only a function of the recombination rate between the loci, but also a function of time (number of generations). Tight physical linkage lowers the probability of forming a recombinant in any given generation, but there is no a-priori reason why such recombinants cannot accumulate over many generations (assuming they are not selected against).

At the population level, it is convenient to think about the association between alleles in two loci in terms of linkage disequilibrium. Linkage disequilibrium is defined as the difference in frequency between non-recombinant and recombinant haplotypes, which for *ROS* and *EL* can be written as:

$$D = (p_{ROS\,el} \times p_{ros\,EL}) - (p_{ROS\,EL} \times p_{ros\,el})$$

Where p refers to the frequency of each subscripted haplotype in the population. Linkage disequilibrium can reach a maximum value $D = 0.25,$ when the two parental haplotypes have an equal frequency (i.e., $p_{ROS\,el} = p_{ros\,EL} = 0.5$) and a minimum value $D = 0$ when all four haplotypes occur at the same frequency (i.e., $p_{ROS\,el} = p_{ros\,EL} = p_{ROS\,EL} = p_{ros\,el} = 0.25$). Given a certain recombination rate between two loci, c (which can be approximated to the genetic distance in cM) and an initial value of linkage disequilibrium, D_0, we can calculate D after T generations using the equation (Hedrick 2011):

$$D_T = (1 - c)^T \times D_0$$

From this equation it becomes immediately evident that the decay of linkage disequilibrium is not only determined by the recombination rate, c, between the two linked loci, but also by the number of generations, T, during which recombination can occur (Figure 6.1). In other words, given sufficient time, any allelic association between linked loci can be broken down by recombination until the population reaches linkage equilibrium ($D = 0$).

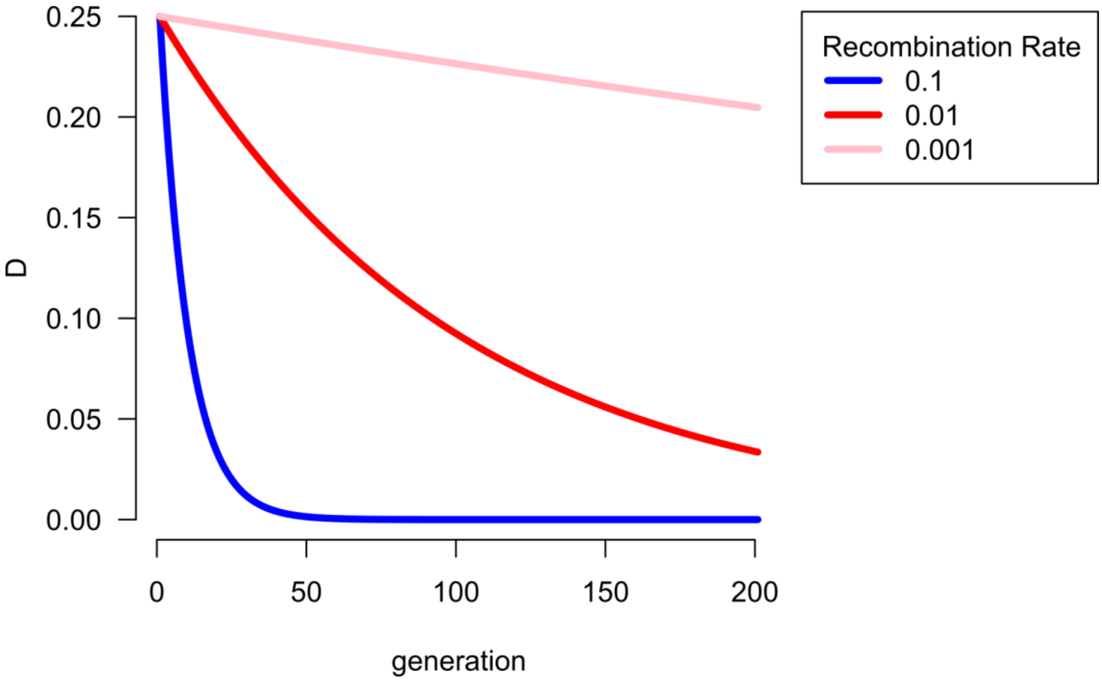


Figure 6.1 – Theoretical decay of linkage disequilibrium over time.
Different recombination rates between two linked loci are considered (coloured lines). In each case, the populations start at the maximum value of $\boldsymbol{D = 0.25}$.

This model for linkage disequilibrium decay assumes that individuals in the population mate randomly and that there is no selection against recombinants. These assumptions may not be true for real populations, but nonetheless the model allows us to intuitively see that the accumulation of recombinants in a population is not only dependent on the recombination rate between two loci, but also depends upon how many opportunities (i.e., generations) there is for recombination to take place.

Considering that tight linkage is not an absolute impediment for the accumulation of recombinants in a population, it is reasonable to ask if recombinant *ROS EL* or *ros el* haplotypes can be found in the hybrid zone population. On the one hand, recombinants might not be found, if the hybrid zone is young (not enough time to accumulate recombinants) and/or recombinant haplotypes are strongly selected against. On the other hand, if selection is not too strong and/or there was enough time of contact between populations, recombinant *ROS-EL* haplotypes might be found segregating in the population. To screen for recombinants between *ROS* and *EL* in the hybrid zone population, it is necessary to determine the genotype of individuals for those loci. So far, the genotype of *ROS* and *EL* could be determined from crossing plants to *A. majus* lines of known genotypes (chapter 4). However, this is not feasible for a large sample size of individuals. Such an approach would involve collecting pollen from wild individuals and use it in crosses to *JI7* and *rosea*^{dorsea}. Not only would this be extremely laborious for thousands of individuals, but technically challenging, as pollen would have to be kept fresh until ready for use, and individuals in the wild would have to be flowering so that pollen was available for collection.

One possibility to determine the *ROS* and *EL* genotypes of individuals from the hybrid zone would be to use the phenotype of the flowers to infer the genotype at the two loci. To see if this would work, it is useful to see what the expected phenotypes are, based on the known genetic interactions between *ROS* and *EL* from controlled crosses (Figure 6.2). In summary, individuals that are homozygous *ros/ros* are expected to be non-magenta, regardless of their genotype in *EL*. Conversely, individuals with at least one dominant *ROS* allele (*ROS/ros* or *ROS/ROS*) are expected to produce pigment, which is either restricted (Eluta phenotype) or spread (magenta phenotype), depending on the presence or absence of a dominant *EL* allele in the background (Figure 6.2). From this prediction, it becomes clear that inferring the *ROS-EL* genotypes from the phenotype is not possible, since different genotypes can produce the same phenotype (for example, *ROS el*/*ros el* and *ROS el*/*ROS el* are both magenta).

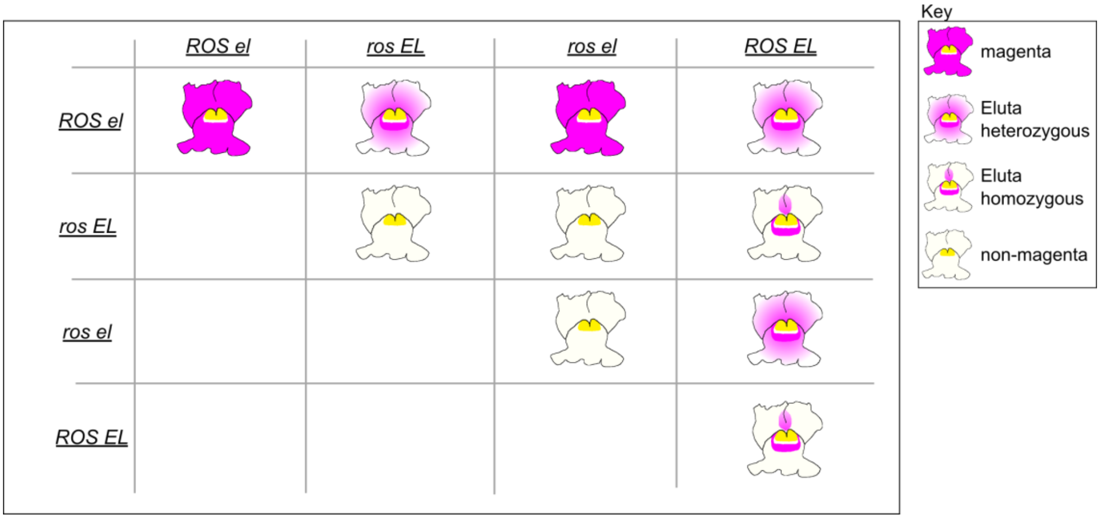


Figure 6.2 – Expected phenotypes from all diploid *ROS-EL* haplotype combinations.
The phenotypes in each case are represented by a drawing that conveys the pigmentation pattern in the face-view of the flowers. In summary, plants with a dominant allele of *ROS* are expected to produce pigment in the flowers. These can have an Eluta phenotype if together with a dominant allele of *EL*, or magenta phenotype if coupled with a recessive *el* allele. In opposition, individuals with recessive alleles of *ros* are always expected to be non-magenta. Notice that because *EL* is semi-dominant, the Eluta phenotype of homozygotes *EL/EL* is more pronounced than that of heterozygotes *EL/el*. Also notice that some genotypes have the same phenotype: for example a *ROS el*/*ROS el* is indistinguishable from a *ROS el*/*ros el* (both magenta) as is a *ros EL*/*ros EL* from a *ros EL*/*ros el* (both non-magenta).

A third solution for the determination of the *ROS-EL* genotypes in the hybrid zone is to use polymorphic molecular markers linked to the two loci, which allow alleles of the two subspecies to be distinguished. In particular, by taking advantage of the results from the mapping experiments, it is known that *ROS1* and *EL-MYB* are part of the *ROS* and *EL* loci, respectively. Therefore, markers in these genes can be used to infer the genotype of *ROS* and *EL* in individuals from the hybrid zone. This approach has successfully been used previously to genotype individuals from the hybrid zone (Whibley 2004; Whibley *et al.* 2006; Elleouet 2012) and was thus the approach chosen for this work.

6.2 Results

6.2.1 Molecular markers in *ROS* and *EL*

To identify the *ROS* and *EL* genotype of plants from the *pseudomajus* x *striatum* hybrid zone, I used polymorphic molecular markers linked to the *ROS1* and *EL-MYB* genes. For each gene, I used two tightly-linked markers: one marker for an indel polymorphism and another marker for a single nucleotide polymorphism (SNP) (Figure 6.3A). *A. m. pseudomajus* and *A. m. striatum* individuals from the flanks of the hybrid zone carry distinct diagnostic molecular alleles for these molecular markers (Whibley 2004; Elleouet 2012). Therefore, these markers can be used to distinguish between *pseudomajus* and *striatum ROS1* and *EL-MYB* alleles.

The indel marker for *ROS1* is located in the promoter region, 94bp upstream of the start codon of this gene. A PCR reaction produces three fragments: two that are typically found in *A. m. pseudomajus* individuals and one that is typical from *A. m. striatum* (Figure 6.3B). The SNP marker is located in the first intron of *ROS1*, only 356bp downstream of the indel marker. The genotype for this marker is obtained by using KASP technology (LGC Genomics) and each of the two alleles from this marker is specific to one subspecies (Figure 6.3B). Therefore, the genotype obtained from one marker can be used to confirm the genotype of the other marker and, together, provide a consensus genotype for the *ROS1* gene (i.e., *ROS1*^{p} or *ros1*^{s}).

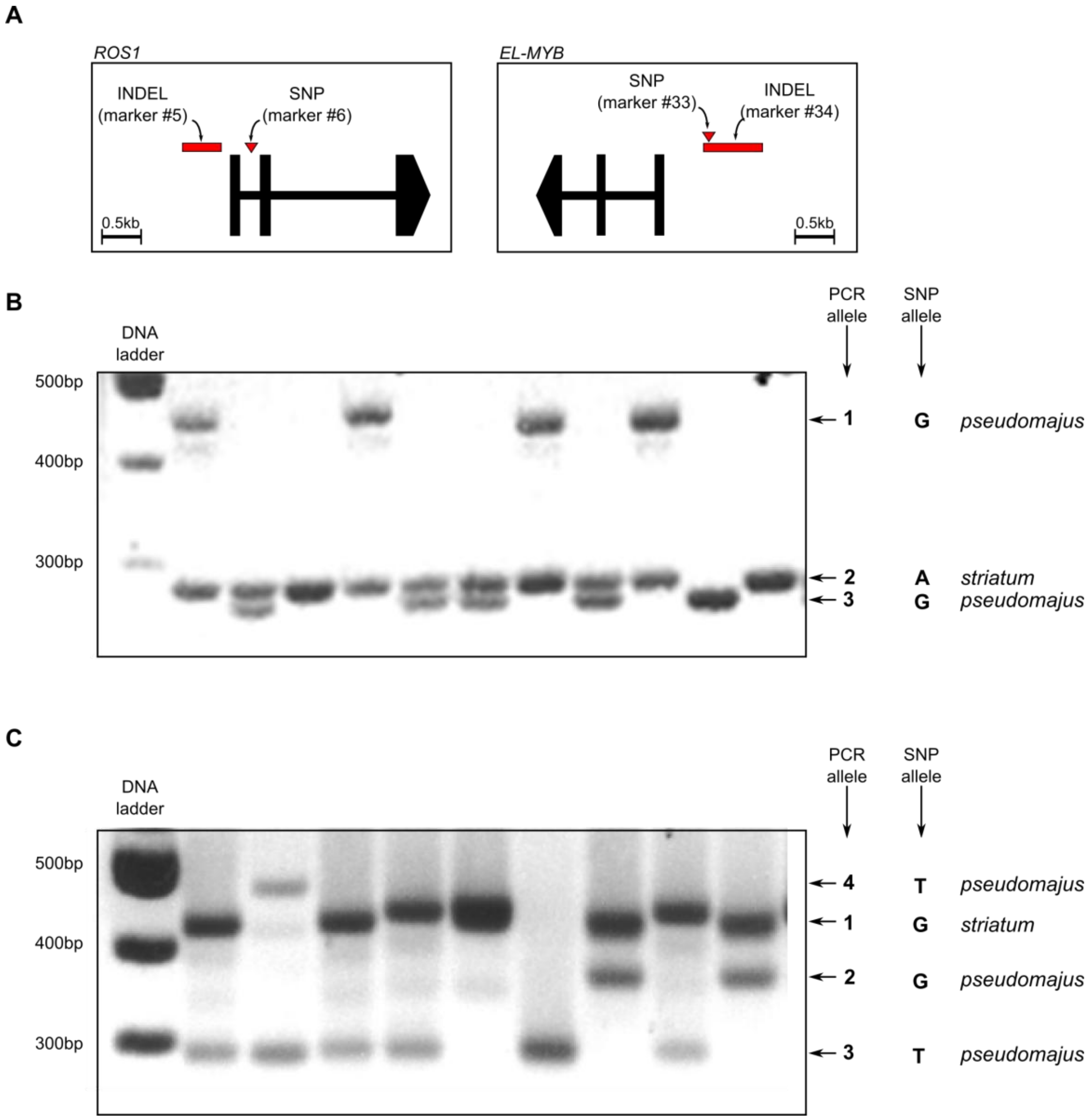


Figure 6.3 – Markers for genotyping *ROS1* and *EL-MYB* in hybrid zone samples.

A) Location of indel and SNP markers in the *ROS1* and *EL-MYB* genes (red boxes and triangles). The marker numbers correspond to Table 2.1. For each gene, the exons are indicated as boxes and the introns as lines connecting them.

B) Example photograph of an agarose gel electrophoresis for the PCR products obtained with the indel marker in *ROS1* (marker #5). Three alleles were distinguished by size and arbitrarily numbered 1-3. The allele from the SNP marker corresponding to each of these PCR fragments is indicated, as is the consensus subspecies allele (*pseudomajus* or *striatum*) that they identify.

C) Same as B) but for the indel marker in *EL-MYB* (marker #34). Four alleles could be distinguished by size and were arbitrarily numbered 1-4. Notice that allele 1 visibly includes differently sized fragments; however, because these cannot be confidently resolved in an agarose electrophoresis, they were scored as a single allele. Also notice that the SNP marker is not fully diagnostic between the two subspecies alleles, but can be used together with the PCR marker to determine a consensus genotype.

The indel marker for *EL-MYB* is located in the promoter region of the gene, being 2287bp upstream of its start codon. This marker is obtained by PCR and produces four fragments: three of them typically correspond to *A. m. pseudomajus* and one of them to *A. m. striatum* (Figure 6.3C). The SNP marker (also obtained by KASP technology) is located within the PCR fragment of the indel marker. This marker is not fully diagnostic between the two subspecies, but has a consistent association with certain alleles of the indel marker (Figure 6.3C). Therefore, the two markers can be used together to determine a consensus genotype for *EL-MYB* (i.e., *el-myb*^{p} or *EL-MYB*^{s}).

Using these markers, haplotypes can be determined by considering the diploid genotype of an individual in *ROS1* and *EL-MYB* (Figure 6.4). In particular, recombinants can be found by finding discrepancies in the genotypes of the two genes. For example, if an individual is determined to be heterozygous *ROS1*^{p}/*ros1*^{s} but homozygous *el-myb*^{p}/*el-myb*^{p}, it suggests that it carries one parental haplotype – *ROS1*^{p} *el-myb*^{p} – and one recombinant haplotype – *ros1*^{s} *el-myb*^{p}. The one caveat to this method is that heterozygous individuals with the two subspecies haplotypes (*ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *EL-MYB*^{s}) cannot be distinguished from heterozygous individuals with two reciprocal recombinant haplotypes (*ROS1*^{p} *EL-MYB*^{s} / *ros1*^{s} *el-myb*^{p}) as both will be heterozygous for the markers in both genes (central panel in Figure 6.4). However, if the frequency of recombinant haplotypes in the population is not very high, the occurrence of the latter genotype should be sufficiently rare to be ignored when calculating haplotype frequencies.

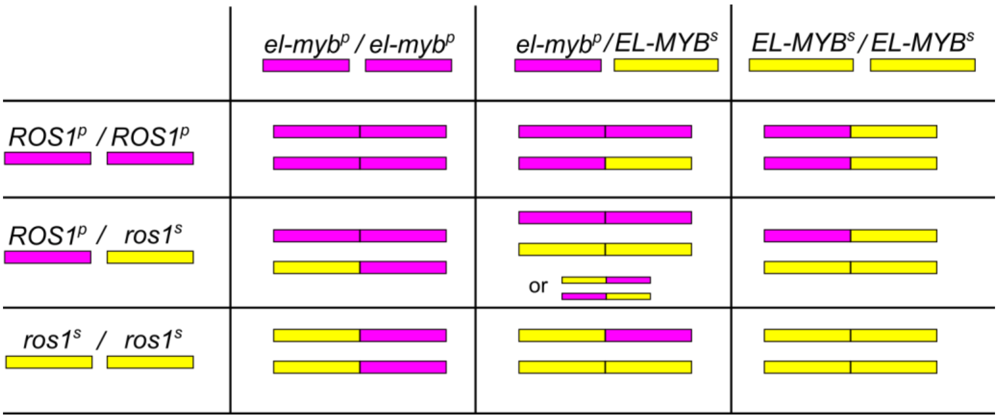


Figure 6.4 – Determination of *ROS-EL* haplotypes from genotypes in *ROS1* and *EL-MYB*. Haplotypes can be determined from the genotypes in each gene. The coloured bars denote the *A. m. pseudomajus* (magenta) and *A. m. striatum* (yellow) alleles in each gene. Notice that in one case (central panel) the haplotypes are ambiguous: if an individual is heterozygous for both genes, it is impossible to distinguish if it carries the two subspecies haplotypes or two reciprocal recombinant haplotypes. Because the latter should be rare in the population, individuals heterozygous for both genes are always assumed to carry non-recombinant haplotypes.

In terms of nomenclature for this chapter, I will refer to the molecularly-determined alleles in each gene as *ROS1*^{p}, *ros1*^{s}, *el-myb*^{p} and *EL-MYB*^{s}. Haplotypes will be denoted using the usual notation, for example *ROS1*^{p} *el-myb*^{p} for an *A. m. pseudomajus* haplotype. It is important to clarify the distinction between a molecular genotype and a functional genotype. In the context of this work, a molecular genotype is determined by the markers located near a gene, whereas the functional genotype is determined by crosses to *A. majus* lines of known genotype (as in Figure 4.3). Therefore, the functional genotype refers to the genetically defined *ROS* and *EL* loci, which may include several genes or functional sites within them (as discussed in the mapping experiments of chapter 4). The molecular genotype is expected to predict the functional genotype but, as shall be seen, this is not always the case. The molecular genotypes will always refer to a gene allele (in this chapter for *ROS1*, *ROS2* and *EL-MYB*), whereas the functional genotypes refer to the general genetic locus (*ROS* and *EL*).

6.2.2 Screening *ROS-EL* recombinants in a natural hybrid zone population

The markers in *ROS1* and *EL-MYB* were used to genotype 2393 individuals sampled from the hybrid zone. The geographic location of these individuals is spread across the population, although the number of individuals genotyped was not similar across the flower colour cline (Table 6.1). Since the main aim of this experiment was to find recombinants between *ROS* and *EL* in the natural population, the sampling effort was concentrated nearer the centre of the cline. This is because of the expectation that recombinants should be commonest where the two parental *ROS-EL* haplotypes occur at similar frequencies, that is, in the "core" of the hybrid zone.

Table 6.1 – Number of genotyped individuals at different distances from centre of the flower colour cline.
The number of genotyped individuals is given within a certain Euclidian distance from the canonical centre of the flower colour cline. From the total number of genotyped individuals, a very high fraction (90-95%) was successfully genotyped at all four markers used in *ROS1* and *EL-MYB*.

Distance from hybrid zone centre (km)	Number of samples (successfully genotyped)
≤ 0.5	1359 (1295)
0.5 to 2	659 (629)
> 2	375 (339)

Out of the 2393 individuals genotyped, 2263 (~95%) were successfully genotyped with all four markers (Figure 6.3). Some individuals had a discrepancy between the expected association of the indel and SNP alleles co-located in either *ROS1* or *EL-MYB*. From a total of 4530 haplotypes (two from each of the 2263 genotyped individuals), 90% were unambiguously determined from the molecular markers as being *pseudomajus*, *striatum* or recombinant. The observed discrepancies between the indel and SNP markers within each gene can be due to several reasons, namely: technical errors during the genotyping; null PCR alleles (cases where the primers do not work well for certain alleles, therefore a homozygous individual for a molecular marker may actually be heterozygous with an allele that was not amplified); rare haplotypes segregating in the population. These cases were not investigated thoroughly and, since the sample size used was large and there was no geographic bias in their occurrence, they were removed from the analysis.

There are four possible haplotypes that can theoretically form: two subspecies haplotypes – *ROS1^p el-myb^p* and *ros1^s EL-MYB^s* – and two recombinant ones – *ROS1^p EL-MYB^s* and *ros1^s el-myb^p*. All these haplotypes were observed in the hybrid zone (Table 6.2). The majority of haplotypes consisted of the two subspecies combinations, but around 5% of them were recombinant haplotypes. These haplotypes do not occur at a homogenous frequency across the geographic range of this population. The parental haplotypes show a gradual change in frequency across the hybrid zone, which is correlated with the change in flower colour (Figure 6.5). The *pseudomajus* haplotype declines in frequency from East to West, as the magenta phenotype becomes rarer. In opposition, the *striatum* haplotype declines in frequency from West to East, as the non-magenta phenotype becomes rarer. The recombinant haplotypes show an increased frequency near the centre of the hybrid zone and a reduced frequency in its flanks (Figure 6.5B). Although the frequency of the recombinant haplotypes is ~5% in the total sample (Table 6.2), their frequency can reach up to 12.5% in local clusters of plants within a 200m radius.

Table 6.2 – Number of recombinant and parental haplotypes observed in the hybrid zone population.

Haplotype	Number observed (percentage of total)
ROS1^p el-myb^p	2047 (50%)
ros1^s EL-MYB^s	1845 (45%)
ROS1^p EL-MYB^s	134 (3%)
ros1^s el-myb^p	67 (2%)

There is a signal of introgression of both recombinant and *pseudomajus* haplotyes into the *A. m. striatum* flank (Figure 6.5B). Considering samples further than 500m East-West from the centre of the flower colour cline, there are 31 in 812 (3.8%) recombinant haplotypes on the *A. m. striatum* side, and only 5 in 508 (1%) recombinant haplotypes on the *A. m. pseudomajus* side ($p < 0.01$, Fisher's exact test). The same trend is observed for the introgression of each parental haplotype: 56 in 812 (6.9%) *pseudomajus* haplotypes occur in the *A. m. striatum* side, versus only 15 in 508 (3%) *striatum* haplotypes being found in the *A. m. pseudomajus* side ($p<0.01$, Fisher's exact test). This result suggests there is a general introgression of non-*striatum* haplotypes to the *A. m. striatum* flank of the hybrid zone. Possible reasons for the introgression of recombinant haplotypes are re-visited in the discussion.

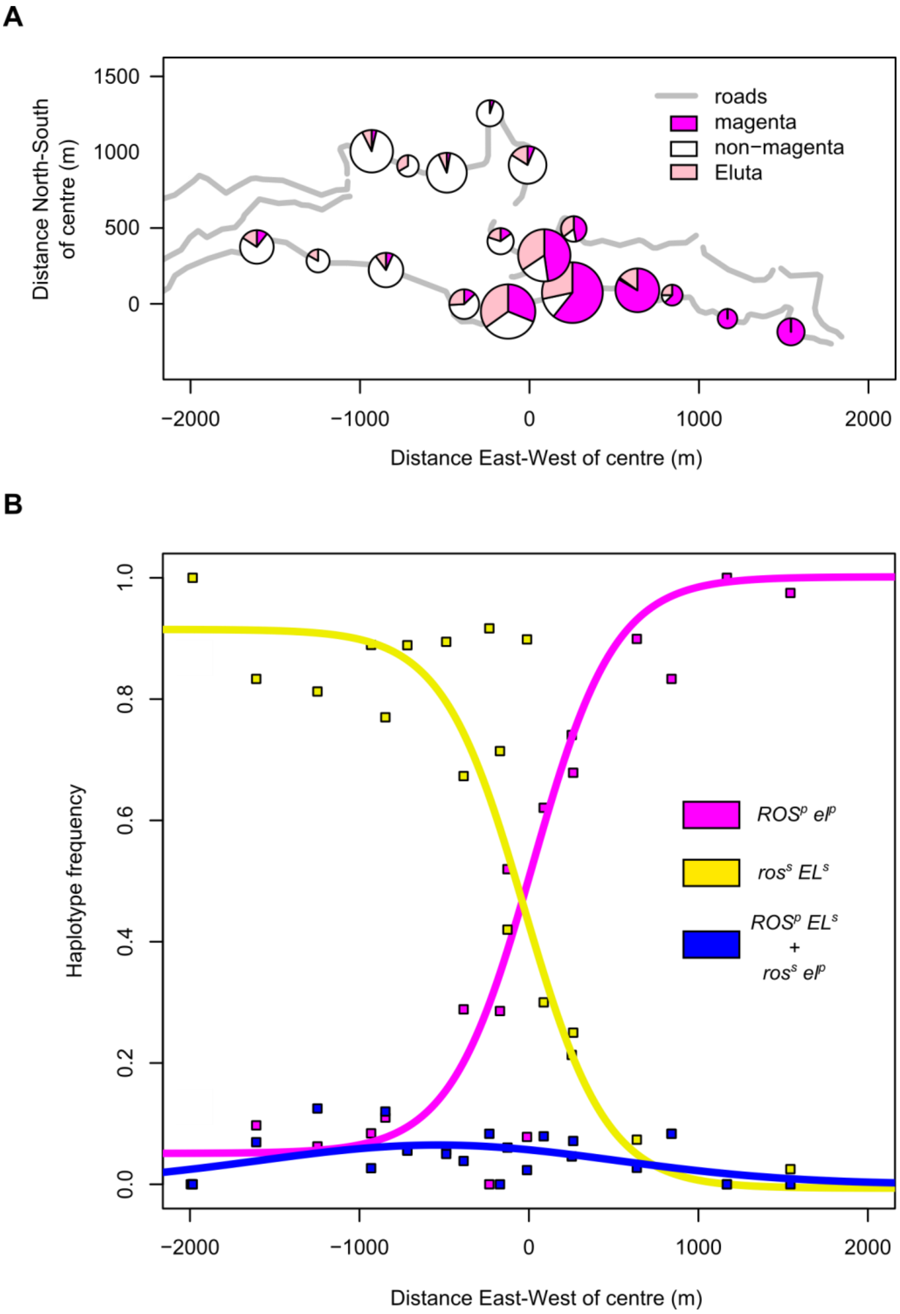


Figure 6.5 – Variation of phenotypes and haplotypes across the hybrid zone.

A) Frequency of phenotypes represented as pie charts. Each pie chart represents a cluster of plants within a 200m radius and is plotted on the geographic location of that cluster's centre. The area of each chart is proportional to the number of samples in the cluster (between 6 and 577 individuals). Each individual's flower was scored in the field for the magenta phenotype and scores were classified as: non-magenta (score < 1), Eluta (1 ≤ score < 3) and magenta (score ≥ 3).

B) Frequency of haplotypes along the West-East direction. Each point is the mean frequency in a cluster of plants within a 200m radius (same clusters as shown in panel A). The lines are a fit to the data, using a 4-parameter sigmoidal function for the parental haplotypes and a Gaussian function for the recombinant haplotypes (section 2.8 in methods).

The value of linkage disequilibrium between *ROS1* and *EL-MYB* can be calculated and used to estimate an age for the hybrid zone, under the assumption that individuals in the population mate randomly (section 6.1). Linkage disequilibrium is expected to be high, since the parental and recombinant haplotypes are represented at very different frequencies across the whole dataset (95% and 5%, respectively). However, because of the unequal distribution of haplotypes across the hybrid zone (Figure 6.5B), estimates of linkage disequilibrium may be over-estimated if considering the flanks of the hybrid zone, where one type of haplotype is more frequent than any other. Therefore, I calculated linkage disequilibrium only for samples within 500m East-West distance from the centre of the flower colour cline. As expected, the *pseudomajus* and *striatum* alleles in each gene are in high linkage disequilibrium in the centre of the hybrid zone: $D = p_{ROS\,el} \;\; p_{ros\,EL} - p_{ROS\,EL} \;\; p_{ros\,el} = 0.22$. This value can be used to estimate an age for the hybrid zone (under the assumption of random mating in the central region of the hybrid zone), by using the equation $D_T = (1-c)^T \times D_0$ (section 6.1). By solving the equation as a function of T, we obtain:

$$T = log_{(1-c)}(D_T/D_0)$$

We can assume that, when the populations met, there were no recombinant haplotypes (only <u>*ROS1^p el-myb^p*</u> and <u>*ros1^s EL-myb^s*</u>), and therefore $D_0 = 0.25$. We can also consider the recombination rate to be $c = 0.005$, based on the map distance between *ROS* and *EL* determined in mapping experiments (0.5cM). By using the currently observed value of $D = 0.22$, we obtain an estimate of $T = 25.5$ generations of hybridization. Assuming that *A. majus* are annual plants (one generation per year) the age of the hybrid zone could be approximated to ~25 years. This estimate assumes that there is random mating in the population and that recombinants are not selected against. If this assumption is not valid, the age of the hybrid zone may be underestimated, as non-random mating would increase the time necessary to break down the *A. m. pseudomajus* and *A. m. striatum ROS-EL* haplotypes. Non-random mating may be due to selection (e.g. if there is assortative mating between individuals of similar phenotypes), or due to non-selective processes. This simple model of LD decay ignores the spatial structure in the hybrid zone (i.e. the fact that allele frequencies change across the cline), which will restrict the region where the recombinants can be formed. Also, individuals that migrate from the flanks (where the "pure" subspecies haplotypes occur) into the centre of the hybrid zone will inflate the levels of LD. Moreover, the same recombinant haplotype might be sampled several times, if it originates from the same recombination event. And finally, despite *Antirrhinum* being mostly annual, there might be a soil seed bank, and thus individuals might originate from several generations ago, potentially biasing the frequency of haplotypes found in any one year of sampling (the dormancy time of *A. majus* seeds is unknown). Despite these known biases, this calculation illustrates that even if there is tight linkage between loci, recombinants can accumulate in relatively little time in a mixed population (with the idealized assumptions of random mating, annual generations and ignoring population structure).

The important result from this experiment is the confirmation that recombinants between the *ROS1* and *EL-MYB* genes are found in the natural population, reaching a mean frequency of 5% in the central region the hybrid zone. This offers the possibility of investigating how the phenotype of the hybrid zone individuals is affected by the particular *ROS-EL* haplotype combinations that they carry.

6.2.3 Genotype-phenotype association in the hybrid zone

To investigate how recombination between *ROS1* and *EL-MYB* affects the phenotype of the individuals in the hybrid zone, I looked at the association between their genotype and the magenta phenotype of the flowers. One assumption from the molecular genotyping is that the molecular markers in *ROS1* and *EL-MYB* are tightly linked to the functional polymorphisms that characterize the *A. m. pseudomajus* and *A. m. striatum ROS* and *EL* loci, respectively. If this is true, then certain phenotypes are expected for different diploid haplotype combinations, based on the interaction between the two loci determined from genetic experiments (Figure 6.6A).

The observed phenotypes for the *ROS1* and *EL-MYB* haplotype combinations observed in the hybrid zone largely fit the expectation from previous genetic experiments (Figure 6.6B). All of the individuals carrying a recombinant haplotype were heterozygous with a non-recombinant haplotype from the subspecies. Because of the genetic epistasis between *ROS* and *EL,* the same recombinant haplotype could result in different phenotypes, depending on which subspecies' haplotype it was heterozygous with (Figure 6.6).

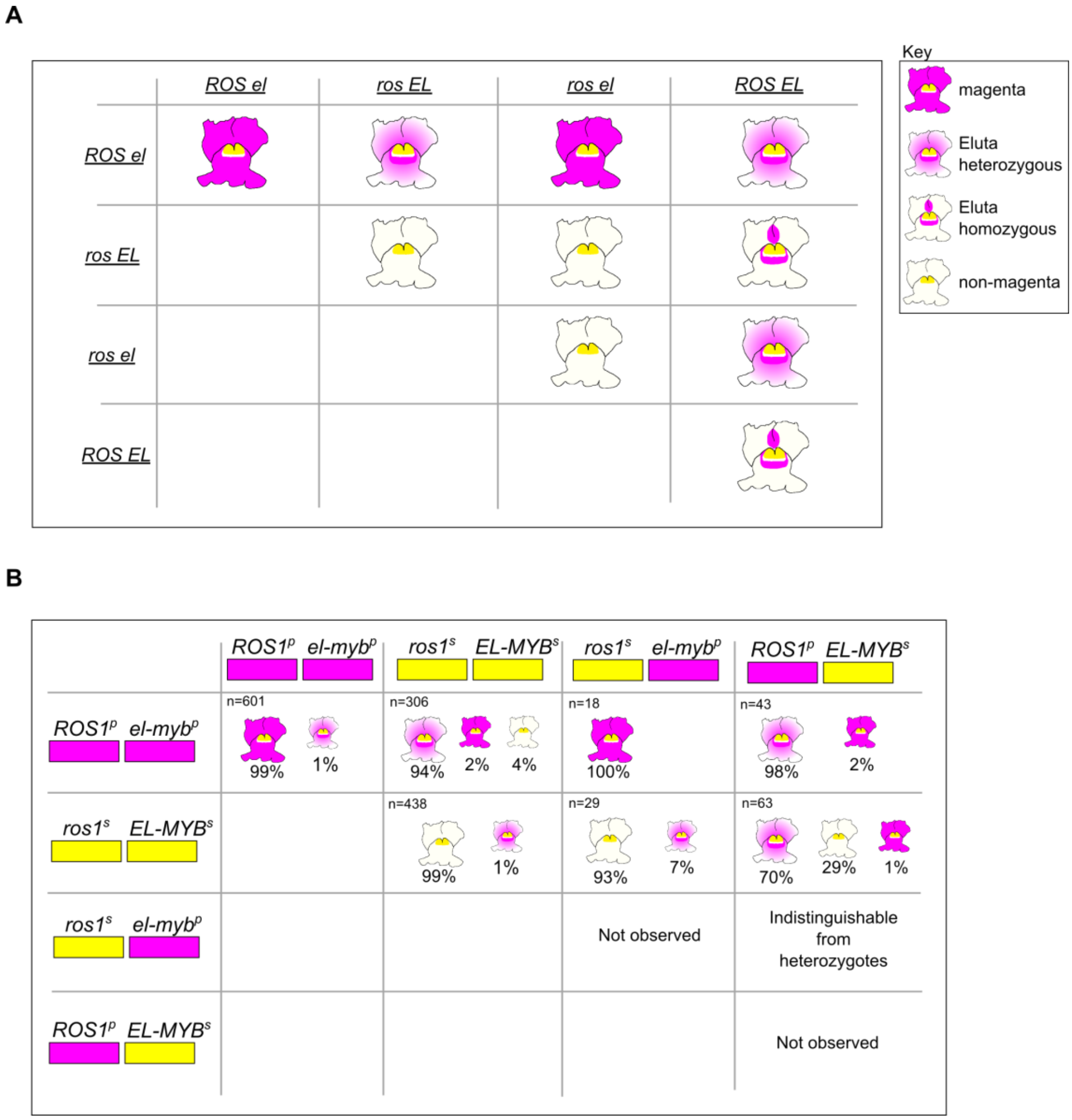


Figure 6.6 – Expected and observed phenotypes of ROS-EL genotypes in the hybrid zone.

A) Expected phenotypes for all diploid *ROS-EL* haplotype combinations (considering functional haplotypes). Same caption as Figure 6.2.

B) Observed phenotypes from several diploid haplotype combinations of *ROS1* and *EL-MYB* (based on molecular markers). Overall, the observed phenotypes fit the expected phenotypes if *ROS1* and *EL-MYB* correspond to the functional *ROS* and *EL* loci, respectively. However, some exceptions are observed (discussed in text). The total number of observations is indicated in each cell, with the percentage of each phenotype indicated. For ease of interpretation, the alleles for each molecular genotype are coloured as boxes (magenta – *pseudomajus*; yellow – *striatum*). Notice that the semi-dominance of *EL* is ignored for the purpose of classifying phenotypes in the hybrid zone due to the high variability of magenta intensity which does not allow confidently distinguishing between homozygous *EL/EL* and heterozygous *EL/el*.

Most individuals with a *ros1*^{s} *el-myb*^{p} haplotype produced magenta flowers if heterozygous with the *pseudomajus* haplotype (*ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *el-myb*^{p}), but non-magenta flowers if heterozygous with a striatum haplotype (*ros1*^{s} *EL-MYB*^{s} / *ros1*^{s} *el-myb*^{p}). Conversely, most individuals with *ROS1*^{p} *EL-MYB*^{s} haplotype produced Eluta flowers, regardless of which subspecies haplotype they were heterozygous with. There were several cases where the observed phenotype did not agree with the expectation for *ROS* and *EL* genotypes. Usually, this occurred in a low proportion of the individuals (1-2%), which may be expected from errors in genotyping or scoring of the phenotype. Other discrepancies may be related to the fact that the molecular markers (in *ROS1* and *EL-MYB*) are not coincident with the functional loci, although tightly linked to them. Therefore, recombination between the molecular markers and the functional polymorphisms in each gene could lead to apparent phenotype-genotype discrepancies (detailed in section 6.2.5).

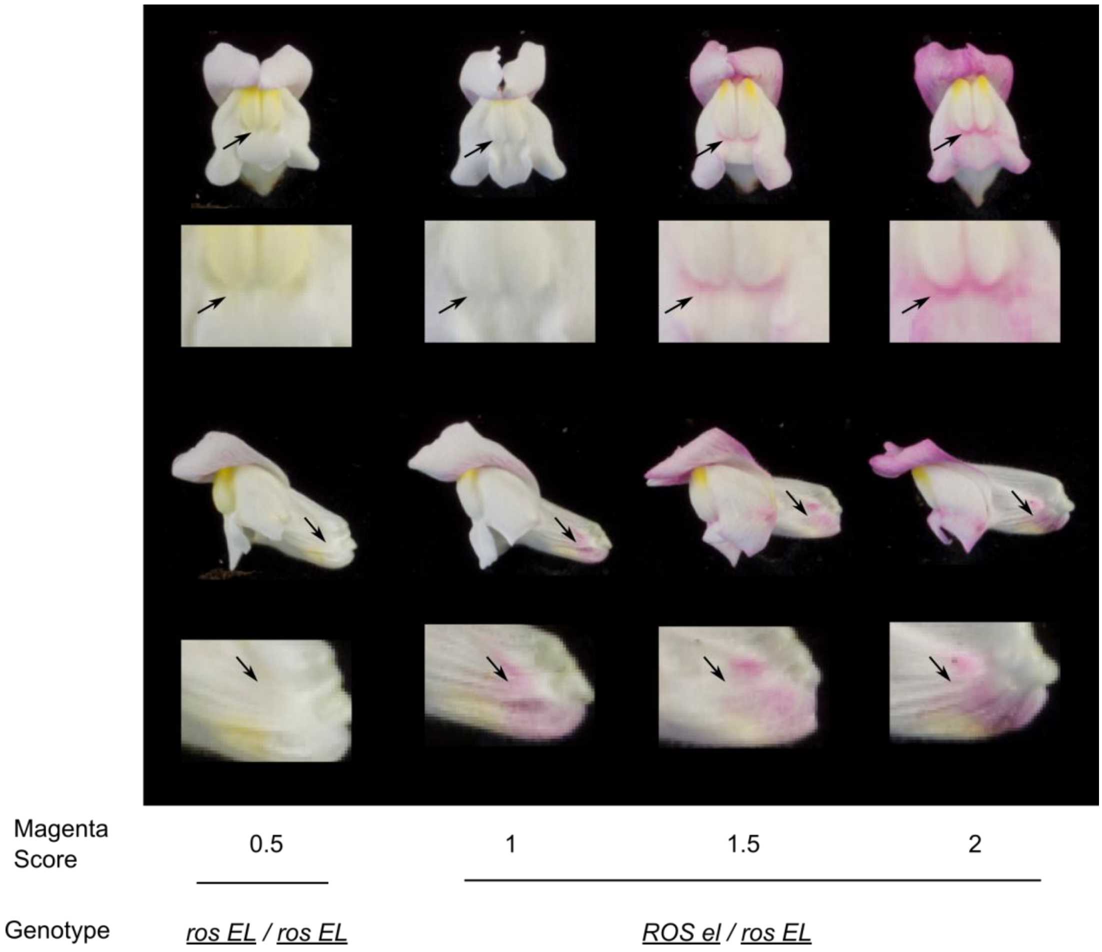


Figure 6.7 – Example of variation in the Eluta phenotype of *ROS el*/*ros EL* heterozygotes from a *pseudomajus* x *striatum* F2.
The photographs represent individuals with an Eluta phenotype (1 ≤ magenta score ≤ 2) and an individual with non-magenta phenotype (magenta score = 0.5) for comparison. Notice that the individual with magenta score 1 does not have visible pigmentation in the flower lobes, but has visible pigmentation in the base of the tube. Due to this subtlety of the Eluta phenotype, some individuals from the hybrid zone population may have been mis-scored as non-magenta (i.e. magenta score < 1). Individuals in photographs are the same as those in Figure 4.5.

There was one case where the phenotype-genotype discrepancy was quite significant: 29% of heterozygotes *ros1*^{s} *EL-MYB*^{s} / *ROS1*^{p} *EL-MYB*^{s}, which are expected to have an Eluta phenotype, had a non-magenta phenotype (Figure 6.6B). This may be explained by the semi-dominance of *EL*, which results in a stronger reduction of pigment in individuals homozygous for a dominant allele in this locus (for example, compare the homozygous and heterozygous line of *ELUTA* in Figure 4.1). If the effect of *EL* is strong enough to overcome the presence of a dominant *ROS* allele in the background, it may result in a non-magenta phenotype, explaining this result. In fact, Whibley (2004) reported that in *pseudomajus* x *striatum* F2s some heterozygotes *ROS1*^{p}/*ros1*^{s} (most of them presumed to be *ROS el*/*ros EL*) had a seemingly non-magenta phenotype. However, upon closer inspection, it turned out that they could be distinguished by the presence of magenta pigment in the base of the flower tube, despite its absence in the flower lobes (Figure 6.7). Therefore, some of the individuals from the hybrid zone may in fact have been wrongly scored as non-magenta (magenta score < 1), by letting the subtle pigmentation in the base of the tube go unnoticed. This is likely to explain the high fraction of non-magenta *ros1*^{s} *EL-MYB*^{s} / *ROS1*^{p} *EL-MYB*^{s} individuals, but also the 4% of cases where heterozygous *ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *EL-myb*^{s} (expected to be Eluta) have that phenotype.

6.2.4 Genetic and phenotypic analysis of naturally occurring recombinants

The analysis of molecular recombinants between *ROS1* and *EL-MYB* in the hybrid zone assumes that the molecular markers correctly identify each subspecies alleles. Moreover, interpretations of the phenotypic consequences of recombination further assume that the molecular markers are tightly linked to the functional *ROS* and *EL* loci (i.e., that the molecular genotype is equivalent to the functional genotype). Although these are reasonable assumptions, it is desirable to confirm the functional genotype for molecularly identified recombinants, particularly considering the phenotype-haplotype discrepancies described above (Figure 6.6).

Therefore, putative recombinant haplotypes identified in the hybrid zone with the molecular markers in *ROS1* and *EL-MYB* were genetically analysed in glasshouse crosses (Figure 6.8). Several individuals sampled in the field season of 2012 (~June) were genotyped with the *ROS1* and *EL-MYB* markers over summer, and putative recombinants (*ROS1*^{p} *EL-MYB*^{s} or *ros1*^{s} *el-myb*^{s}) identified. In September, as the flowering season reached its end and the seed capsules had ripened, the putative recombinants were located back in the field, aided by their GPS coordinates and identification tag. One capsule from each of 135 putative recombinants was collected and their seeds grown in the glasshouse. Some capsules had very low numbers or unviable seeds, and thus progeny was successfully grown for only 39 of those capsules (hybrid zone progeny in Figure 6.8). A total of 385 individuals were grown, with the numbers from each capsule varying between 1 to 13 individuals.

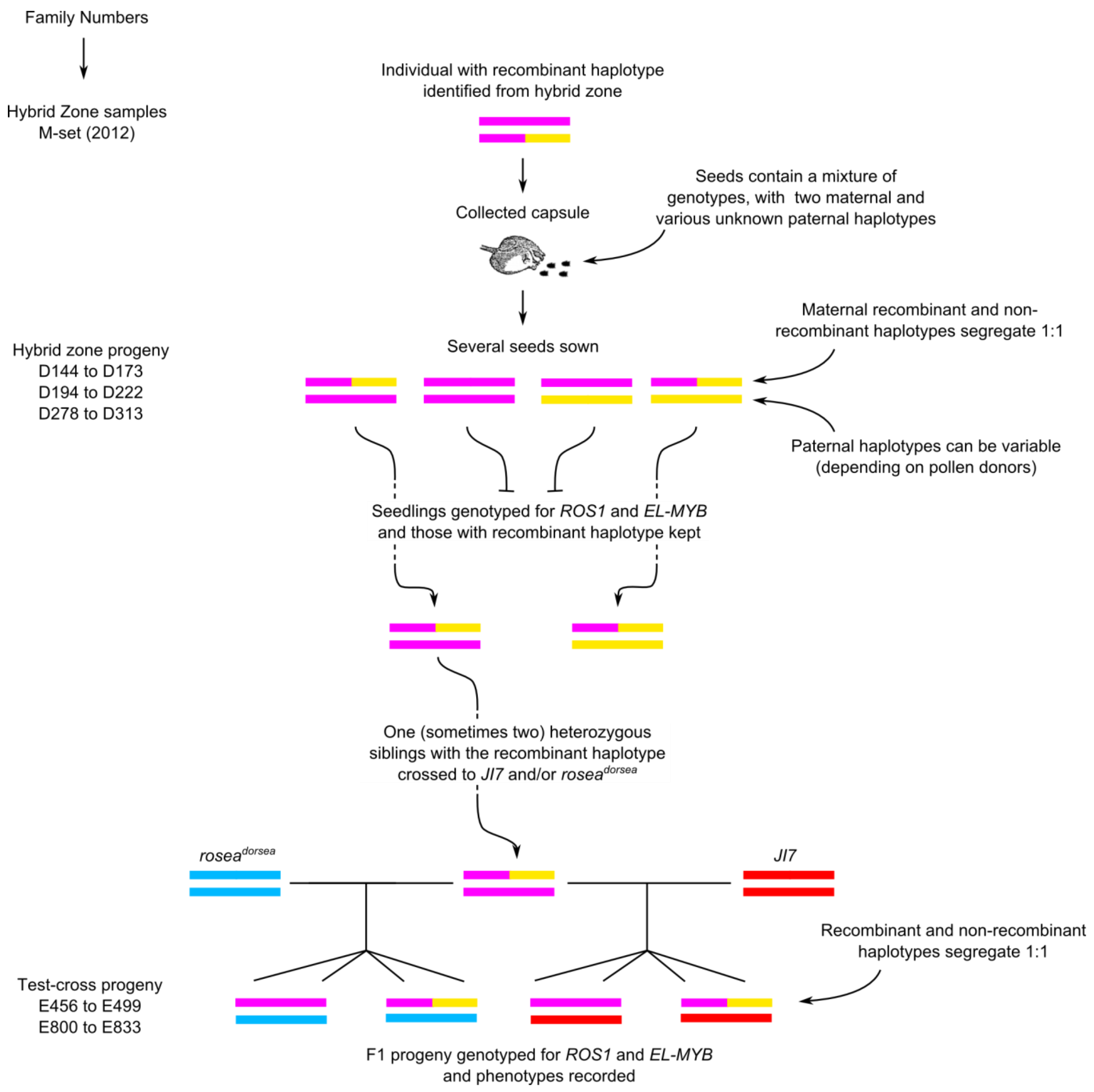

Family Numbers
Hybrid Zone samples
M-set (2012)
Individual with recombinant haplotype identified from hybrid zone
Collected capsule
Seeds contain a mixture of genotypes, with two maternal and various unknown paternal haplotypes
Hybrid zone progeny
D144 to D173
D194 to D222
D278 to D313
Several seeds sown
Maternal recombinant and non-recombinant haplotypes segregate 1:1
Paternal haplotypes can be variable (depending on pollen donors)
Seedlings genotyped for ROS1 and EL-MYB and those with recombinant haplotype kept
One (sometimes two) heterozygous siblings with the recombinant haplotype crossed to JI7 and/or rosea^dorsea
rosea^dorsea
JI7
Recombinant and non-recombinant haplotypes segregate 1:1
Test-cross progeny
E456 to E499
E800 to E833
F1 progeny genotyped for ROS1 and EL-MYB and phenotypes recorded

Figure 6.8 – Schematic of strategy to genetically analyse recombinants obtained from the hybrid zone.

In summary, several recombinants were identified in the hybrid zone and their progeny grown in the glasshouse. Individuals carrying a recombinant haplotype were crossed to stocks of known *ROS-EL* genotype (*JI7* and *rosea^{dorsea}*). Because of segregation of recombinant and non-recombinant haplotypes at several stages of this experiment, plants had to be genotyped for markers in *ROS1* and *EL-MYB* (markers in Figure 6.3). Haplotypes are represented as coloured lines and are given as an example only. The family numbers for each generation are given for reference.

The individuals identified in the field were all heterozygous for a recombinant and a non-recombinant (*pseudomajus* or *striatum*) haplotype. Therefore, the hybrid zone progeny grown from each capsule segregated for the two maternal haplotypes and unknown haplotypes from the paternal pollen donor(s) (wild *A. majus* is self-incompatible and pollinated by various large bee species). Thus, the progeny from each capsule had to be genotyped again with the *ROS1* and *EL-MYB* markers. This allowed individuals that carried the maternal recombinant haplotype together with a non-recombinant paternal haplotype to be identified (these could vary between siblings, as each pistil can receive pollen from several donors). From each family, one or two of the individuals carrying the recombinant haplotype were crossed to *JI7* and/or *rosea*^{dorsea} lines, which have known *ROS* and *EL* genotypes (test-cross progeny in Figure 6.8). The phenotype of these F1 progenies can be used to identify the functional genotype at *ROS* and *EL* for each sampled haplotype (as Figure 4.3). Because the plants grown from wild capsules were heterozygous for recombinant and non-recombinant haplotypes, there was 1:1 segregation of those two haplotypes in the F1s of the crosses to *JI7* and *rosea*^{dorsea}. Therefore, several individuals from each F1 were grown and genotyped for the *ROS1* and *EL-MYB* markers (in most cases 4 individuals were grown, but sometimes less than that successfully germinated). A total of 96 unique haplotypes were successfully genotyped for *ROS1* and *EL-MYB* and crossed to *JI7* and/or *rosea*^{dorsea} (Table 6.3; full list of haplotypes attached in supplementary "*hz_haplotypes.xlsx*"). These haplotypes are thus characterized both by a molecular genotype (at *ROS1* and *EL-MYB*) and a functional genotype (for *ROS* and *EL*). I will onwards refer to the individuals of the final generation of this genetic analysis as the "test-cross" progeny (Figure 6.8).

Table 6.3 – Number of haplotypes sampled from the hybrid zone and test-crossed.
The full table of haplotypes is attached in "*hz_haplotypes.xlsx*".

Haplotype	**F1 x *JI7* and F1 x *rosea*^{dorsea}**	**F1 x *rosea*^{dorsea} only**	**F1 x *JI7* only**
***ROS1*^{p} *el-myb*^{p}**	15	10	13
***ros1*^{s} *EL-MYB*^{s}**	9	7	9
***ROS1*^{p} *EL-MYB*^{s}**	7	9	11
***ros1*^{s} *el-myb*^{p}**	3	2	1

The general expectation for the test-cross progeny is that individuals with the *pseudomajus ROS1*^{p} allele produce magenta pigment in the flowers, whereas those with the *striatum ros1*^{s} allele do not. Moreover, those individuals with the *pseudomajus el-myb*^{p} allele should not result in restricted magenta pigment in the flowers, whereas those with the *striatum EL-MYB*^{s} allele should (Eluta phenotype). I note that the magenta pigmentation in the flowers can sometimes occur specifically overlapping the veins of the petals. This aspect of the phenotype is not considered when scoring phenotypes in the hybrid zone individuals, and will be ignored for now (discussed in section 6.2.6). Therefore, a non-magenta phenotype does not include the vein-specific pigmentation that can often be observed in some flowers (for example, the flower in Figure 6.9e is considered non-magenta, even though pigmentation overlapping the veins of the dorsal petals can be observed).

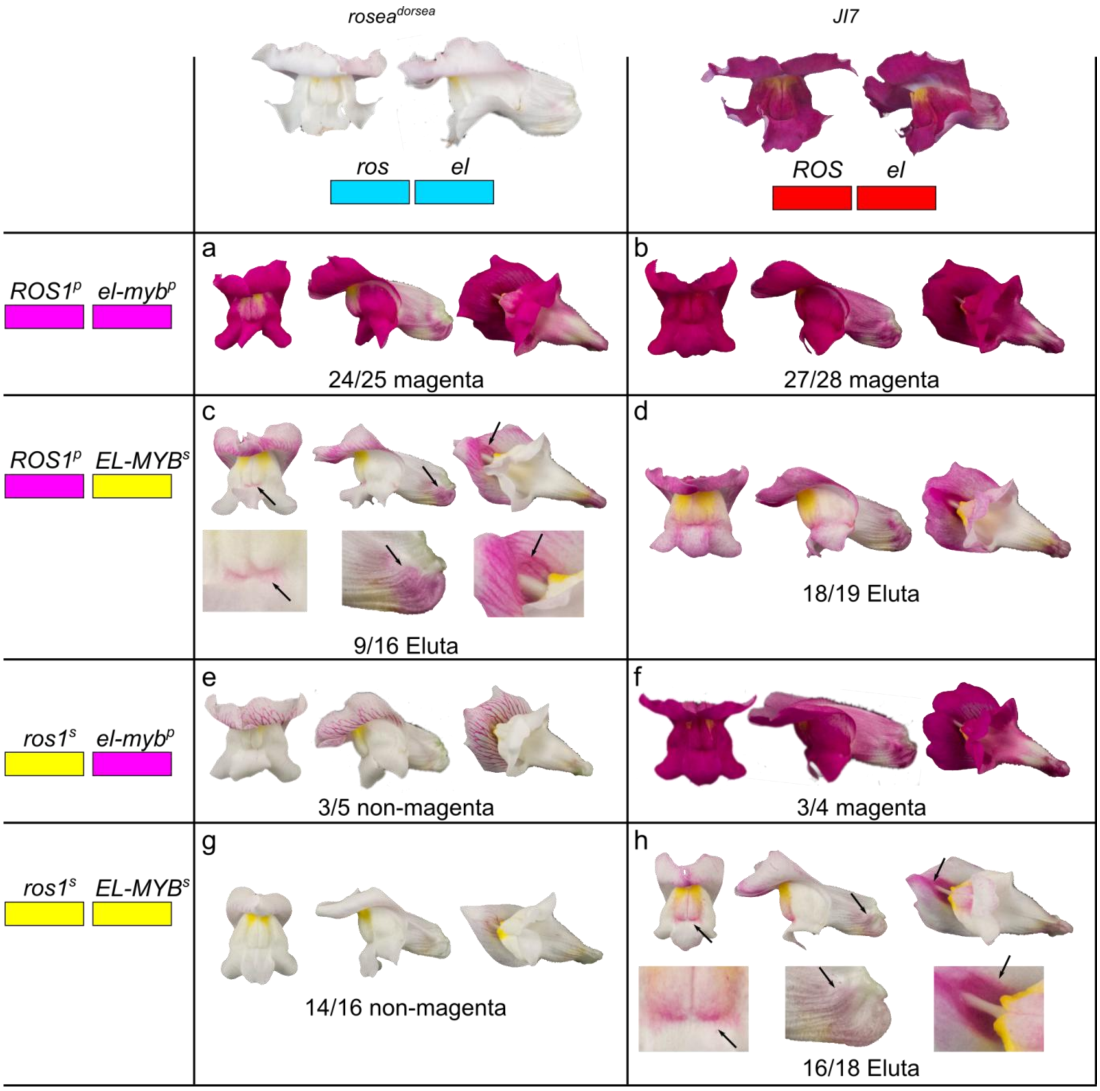


Figure 6.9 – F1 phenotype of test-crosses between individuals sampled from the hybrid zone to *rosea*^{dorsea} and *JI7*.

The functional genotype in each locus can be determined from the phenotype of the test-cross progenies. For each F1, a representative photograph of the most commonly observed phenotype is shown. The number of those phenotypes is indicated in each cell in relation to the total number of haplotypes analysed in each case. Notice that exceptions were observed for every case (discussed in section 6.2.5). The molecular haplotypes are represented with coloured boxes to indicate the origin of each gene's allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea*^{dorsea}. Each cell is referenced with a letter (a to h) for citation in the text. In cases where the pigmentation is difficult to see on the photograph, the pigmented regions of the flower are emphasized by arrows and shown in greater detail below the photograph of the whole flower.

In most cases, the phenotype of the test-cross progeny fitted the prediction from the molecular genotype (Figure 6.9). In particular, individuals with a *ROS1*^{p} *el-myb*^{p} haplotype had magenta flowers in either *JI7* or *rosea*^{dorsea} test-crosses, whereas individuals carrying a *ROS1*^{p} *EL-MYB*^{s} haplotype produced some pigment in the flowers, but restricted with the Eluta pattern (Figure 6.6 a-d). Conversely, the test-cross progeny carrying a *ros1*^{s} *el-myb*^{p} haplotype produced non-magenta flowers, being distinguished from the parental *ros1*^{s} *EL-MYB*^{s} haplotype because they did not confer the Eluta pattern when heterozygous with the *JI7* haplotype (Figure 6.6 e-h). This provides genetic confirmation that most haplotypes identified as molecular recombinants for *ROS1* and *EL-MYB* in the hybrid zone are indeed functional recombinants between *ROS* and *EL*.

6.2.5 Exceptional phenotypes

Several of the observed phenotypes in the test-cross progeny did not match the expected phenotype (Figure 6.9), and thus require an explanation. Because *ROS* includes multiple loci influencing flower colour, some of the discrepancies in the phenotypes could be due to recombination within the *ROS* region. Namely, recombination points between *ROS1* and *ROS2* affect the pigmentation in the flowers, as revealed by the analysis of the rosy phenotype in the mapping experiments of chapter 4 (see Figure 4.12).

To assess the contribution of *ROS2* to the phenotype of the test-cross progeny, I developed a marker in this gene that distinguishes between *pseudomajus* and *striatum* alleles. The marker is located in the third exon of *ROS2* and consists of a

SNP (Figure 6.10). Because of this extra marker, I will need to denote genotypes for three loci, which I will do as *ROS1 ROS2 EL-MYB* (with the usual superscript referring to the allele's origin). This is a simplified notation, and is based on only one marker for *ROS2* located in the third exon of this gene, which is 16.4kb downstream of the other two exons (Figure 6.10). Therefore, recombination points between the markers in *ROS1* (in its promoter and first intron) and the marker in *ROS2* (in its third exon) cannot be resolved.

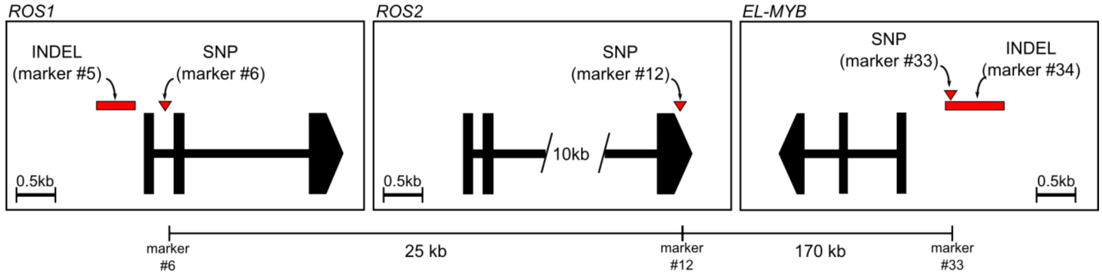


Figure 6.10 – Molecular markers in *ROS1, ROS2* and *EL-MYB* genes used for genotyping haplotypes sampled from the hybrid zone.
The marker position in relation to each gene's coding sequence is given. For each gene, the exons are indicated as boxes and the introns as lines connecting them. The distance between consecutive markers in each gene is given for reference. Marker numbers correspond to Table 2.1.

6.2.5.1 *Incomplete dominance in pseudomajus ROS alleles*

Alleles in the *ROS* locus are considered to be fully dominant/recessive over each other (I am ignoring recombination between *ROS1* and *ROS2*). For example, the F1 progeny of a cross between *JI7* (ROS^{7}/ROS^{7}) and *rosea*^{dorsea} (ros^{dor}/ros^{dor}) produces magenta flowers indistinguishable from *JI7*. However, the analysis of phenotypes in the test-cross progenies suggests that this is not the case for *pseudomajus* ROS^{p} alleles. If the ROS^{p} allele was fully dominant, then the F1 phenotype of a cross to *rosea*^{dorsea} (recessive *ros*) or to *JI7* (dominant *ROS*) should be identical. However, all F1 crosses to *rosea*^{dorsea} produced plants with a weaker intensity of magenta than crosses to *JI7* (Figure 6.11). This suggests that the *ROS*^{p} allele is either haplo-insufficient (that is, two copies are necessary for the darker pigmentation) or that the allele itself is a less strong promoter of anthocyanin production compared to the *JI7* allele.

This observation can be relevant in the context of the phenotypes observed in the hybrid zone. If homozygous *ROS*^{p}/*ROS*^{p} individuals produce more pigment than heterozygous *ROS*^{p}/*ros*^{s}, then certain diploid genotypes may result in different phenotypic outcomes from those predicted (Figure 6.6A). For example, if the production of pigment is limited in heterozygous individuals, then a genotype *ROS*^{p} *el*^{p} / *ros*^{s} *EL*^{s} may result in non-magenta phenotype, if the pigment reduction conferred by *EL*^{s} is sufficiently strong to overcome the contribution of one *ROS*^{p} allele. Indeed, ~4% of heterozygous *ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *EL-MYB*^{s} individuals in the hybrid zone have a non-magenta phenotype (Figure 6.6B), which fits with this hypothesis.

I will maintain the notion of dominant and recessive alleles for *ROS*; however, this result suggests that *ROS* may in fact behave as a semi-dominant locus with regards to the intensity of the magenta pigment.

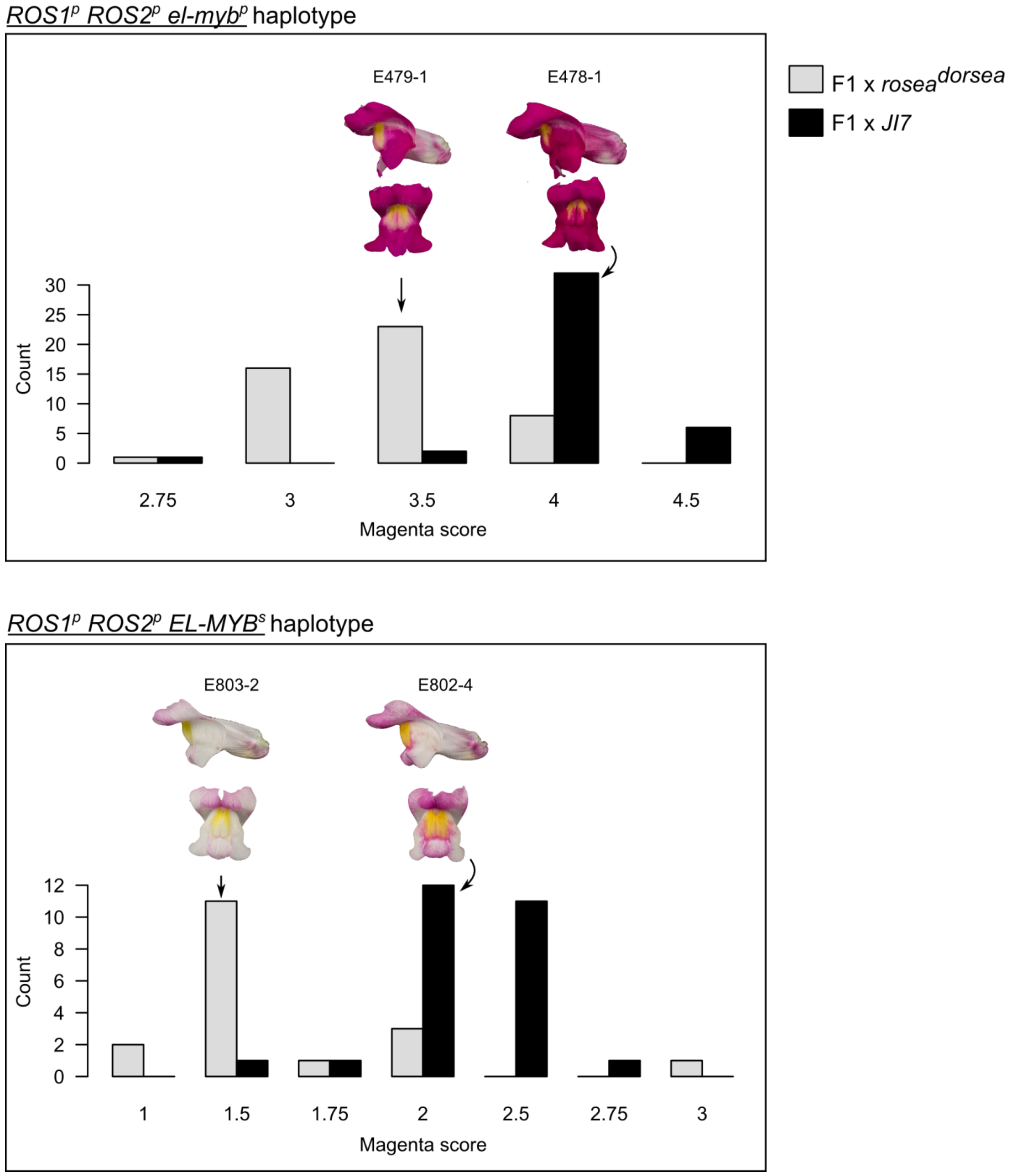


Figure 6.11 – Variation of magenta phenotype intensity related to dominant *ROS* alleles. The two panels show barplots with the count of each phenotypic score for two haplotypes with a dominant allele of $ROS1^{p}$. In each barplot the magenta scores are compared for individuals heterozygous with the $rosea^{dorsea}$ ($ros^{dor}\ el^{dor}$) or the *JI7* ($ROS^{7}\ el^{7}$) haplotypes. Notice that, in both panels, individuals ROS^{p}/ros^{dor} have lower magenta score than those ROS^{p}/ROS^{7}. Representative photographs of flowers with the modal score are shown (the ID of individuals in the photographs are given for reference).

6.2.5.2 Major reduction of anthocynin intensity by putative locus unlinked to ROS-EL

Most test-cross progeny with a parental *A. m. pseudomajus* haplotype (*ROS1*^{p} *ROS2*^{p} *el-myb*^{p}) result in a magenta phenotype when either crossed to *JI7* or *rosea*^{dorsea} (Figure 6.9 a-b). However, there was one exception where the test-cross progeny of this haplotype resulted in significantly paler flowers than in other similar crosses (Figure 6.12 a-b). The pattern of pigmentation of this haplotype crossed to *rosea*^{dorsea} was characterized by a pale tube and pale central region of the lobes, resembling some of the rosy recombinants and the *rosea*^{colorata} line (see Figure 4.12). This similarity of phenotype could imply that rearrangements in the *ROS1-ROS2* region are involved in producing this unexpected phenotype. However, this haplotype bears *pseudomajus* alleles in all used markers (*ROS1*^{p} *ROS2*^{p} *el-myb*^{p}), so there are no detectable recombination points to explain this result.

Furthermore, the phenotype in this test-cross had an inhibitory effect on the pigment, since the cross to the *JI7* haplotype did not result in the usual intensity of magenta in the flowers (compare Figure 6.9b with Figure 6.12b). This effect is distinct from that of rosy or *rosea*^{colorata} haplotypes, both producing normal magenta progeny when crossed to *JI7*, suggesting the involvement of other genetic elements that inhibit the pigment in these flowers.

A similar reduction of pigmentation was observed in the test-cross F1 of a distinct haplotype, bearing the allelic arrangement *ros1*^{s} *ros2*^{s} *el-myb*^{p} (haplotype #69 in Figure 6.12). The test-cross progeny with *JI7* with such a haplotype is expected to result in magenta phenotype (e.g. Figure 6.9f). However, the test-cross F1 of haplotype #69 x *JI7* resulted in variable pigment intensity between 3 siblings (Figure 6.12d). One of these had a phenotype resembling that of haplotype #26 x *JI7* (Figure 6.12 b and d). Despite the small sample size, the segregation of phenotype intensity in the F1 of haplotype #69 x *JI7* suggests that a locus (or multiple loci) unlinked to *ROS-EL* may be responsible for inhibiting the pigment conferred by *ROS* in the flowers. It should be noted that, despite the segregation of magenta intensity in these F1 siblings, all of them are still clearly lighter than flowers seen in other similar F1 crosses (compare Figure 6.12d with Figure 6.9f).

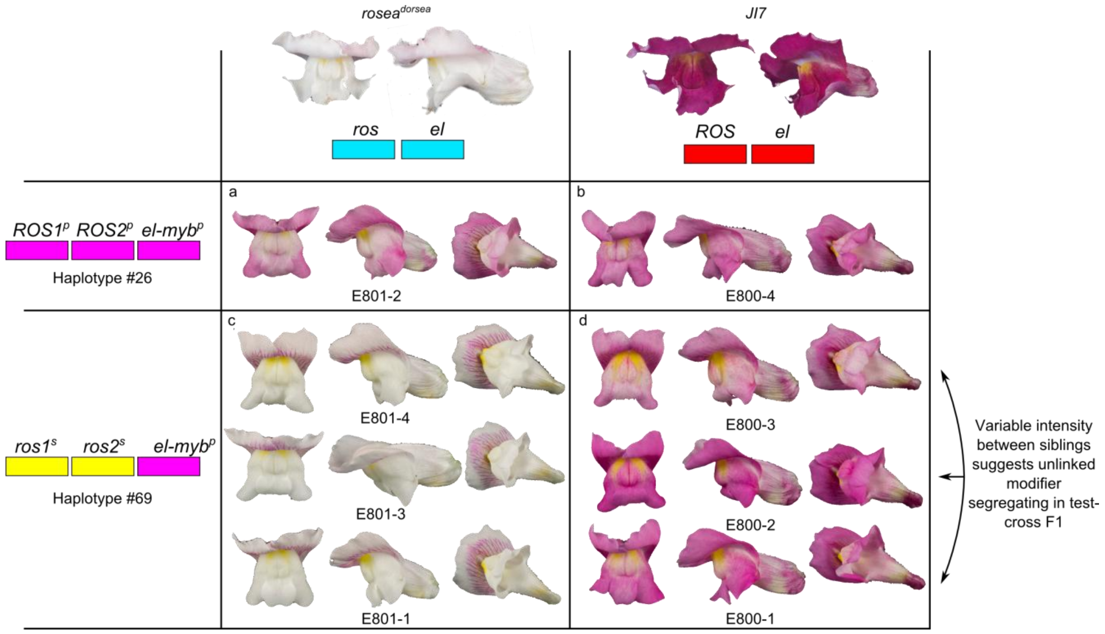

rosea^{dorsea}
JI7
ros
el
ROS
el
a
b
ROS1^p ROS2^p el-myb^p
Haplotype #26
E801-2
E800-4
c
d
ros1^s ros2^s el-myb^p
Haplotype #69
E801-4
E801-3
E801-1
E800-3
E800-2
E800-1
Variable intensity between siblings suggests unlinked modifier segregating in test-cross F1

Figure 6.12 – Reduced pigmentation obtained in test-cross progenies caused by a putative locus (or loci) unlinked to *ROS-EL*.

The photographs in each cell correspond to siblings from each F1 test-cross. Only one plant was produced for the test-crosses involving haplotype #26, whereas three individuals were obtained for each test-cross with haplotype #69. The haplotype numbers correspond to the attached file "*hz_haplotypes.xlsx*". The haplotypes for each gene are represented with coloured boxes to indicate the origin of the allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea^{dorsea}*.

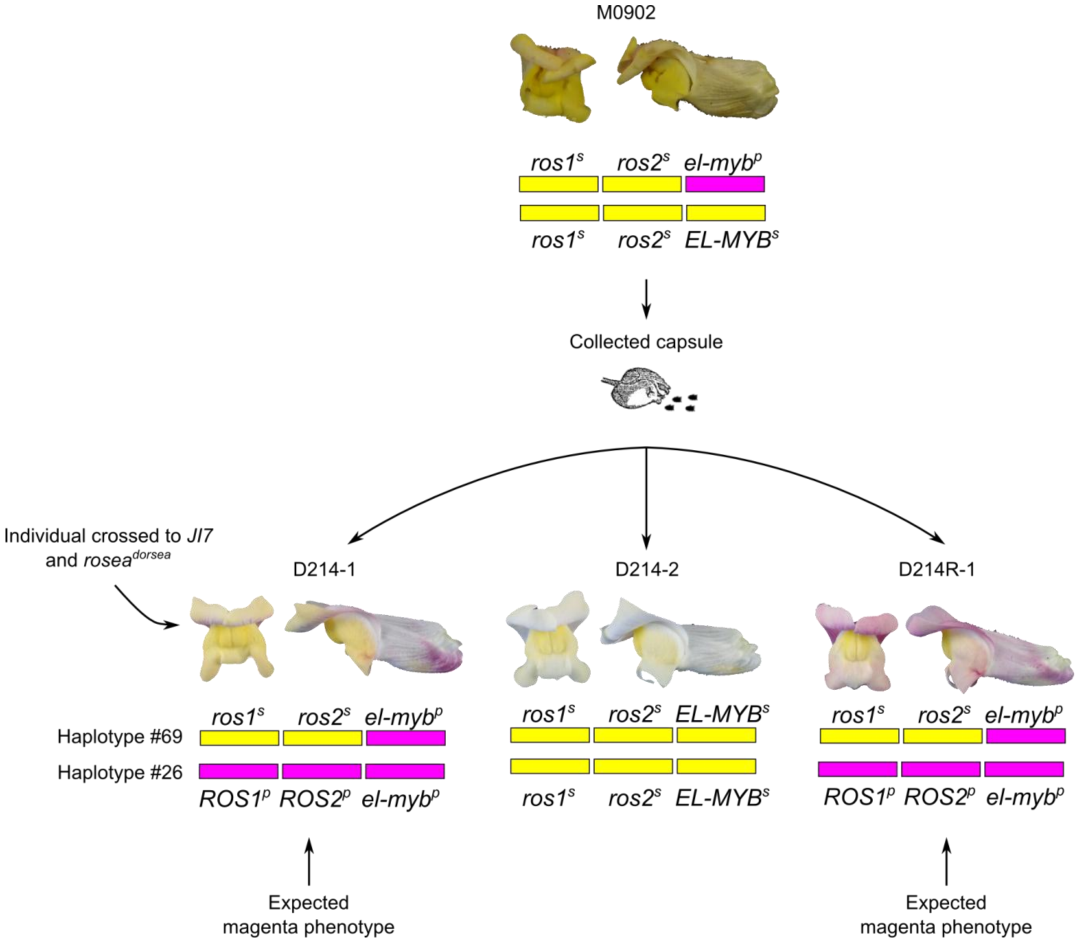

M0902
$ros1^{s}$ $ros2^{s}$ $el\text{-}myb^{p}$
$ros1^{s}$ $ros2^{s}$ $EL\text{-}MYB^{s}$
Collected capsule
Individual crossed to *JI7* and $rosea^{dorsea}$
D214-1
D214-2
D214R-1
Haplotype #69
Haplotype #26
$ros1^{s}$ $ros2^{s}$ $el\text{-}myb^{p}$
$ROS1^{p}$ $ROS2^{p}$ $el\text{-}myb^{p}$
$ros1^{s}$ $ros2^{s}$ $EL\text{-}MYB^{s}$
$ros1^{s}$ $ros2^{s}$ $EL\text{-}MYB^{s}$
$ros1^{s}$ $ros2^{s}$ $el\text{-}myb^{p}$
$ROS1^{p}$ $ROS2^{p}$ $el\text{-}myb^{p}$
Expected magenta phenotype
Expected magenta phenotype

Figure 6.13 – Pedigree of haplotypes showing a novel reduced magenta pigment phenotype.

Three individuals grown from a wild capsule segregated for very pale magenta and non-magenta phenotypes. The pale magenta individuals were expected to have a full magenta phenotype (*$ROS1^{p}$ $el\text{-}myb^{p}$/$ros1^{s}$ $el\text{-}myb^{p}$*). One of these individuals (D214-1) was crossed to *JI7* and *$rosea^{dorsea}$* and the pale magenta phenotype appeared again in the test-cross F1 individuals carrying either of the *ROS-EL* haplotypes (see Figure 6.12).

Further evidence that this effect is due to an element unlinked to *ROS-EL* is the fact that the two haplotypes (#26 and #69) are actually derived from the same parent (Figure 6.13). Therefore, the most parsimonious explanation for the reduced pigment seen in these test-cross progenies is that the parent with these two haplotypes (individual D214-1; Figure 6.13) carried a dominant inhibitory allele in a locus unlinked to *ROS-EL*. This putative inhibitory locus would thus have been transmitted to the progeny independently of the segregation of the *ROS-EL* haplotypes.

Mapping of modifiers of floral pigment other than *ROS-EL* has never been attempted for alleles segregating in the *pseudomajus* x *striatum* hybrid zone. However, this preliminary analysis suggests that they may exist and have a visible impact in the phenotype of the flowers. The test-cross lines produced from this genetic analysis allow this aspect of the phenotype to be investigated further, by providing a genetic resource for mapping putative QTLs associated with flower colour.

6.2.5.3 Influence of elements downstream of ROS1 in floral pigmentation

The test-cross progeny of a recombinant haplotype *ROS EL* should result in flowers with an Eluta phenotype, which is visible in several test-cross F1s involving the molecular haplotype *ROS1^{p} EL-MYB^{s}* (Figure 6.9c-d). The intensity of the pigment is variable depending on the cross being with *JI7* or *rosea^{dorsea}* (possibly due to semi-dominance of *ROS* alleles, as discussed in section 6.2.5.1), but the Eluta phenotype is always evident by the restricted pigmentation to the central part of the lobes and base of the tube.

However, 7 out of 16 *ROS1^{p} EL-MYB^{s}* recombinant haplotypes, when crossed to *rosea^{dorsea}*, had more reduced pigmentation in the flowers than expected (magenta score ≤ 1; Figure 6.14a). Although the junction of the dorsal petals was pigmented as in other Eluta flowers, the pigmentation in the central region of the lobes and base of the tube was mostly absent (compare arrows in Figure 6.14 a and c). All these 7 cases had a *striatum* genotype for the *ROS2* marker, therefore being *ROS1*^{p} *ros2*^{s} *EL-MYB*^{S}. Conversely, the recombinant cases where the Eluta pattern is more evident all carried the *pseudomajus* allele at the *ROS2* marker (*ROS1*^{p} *ROS2*^{p} *EL-MYB*^{S}).

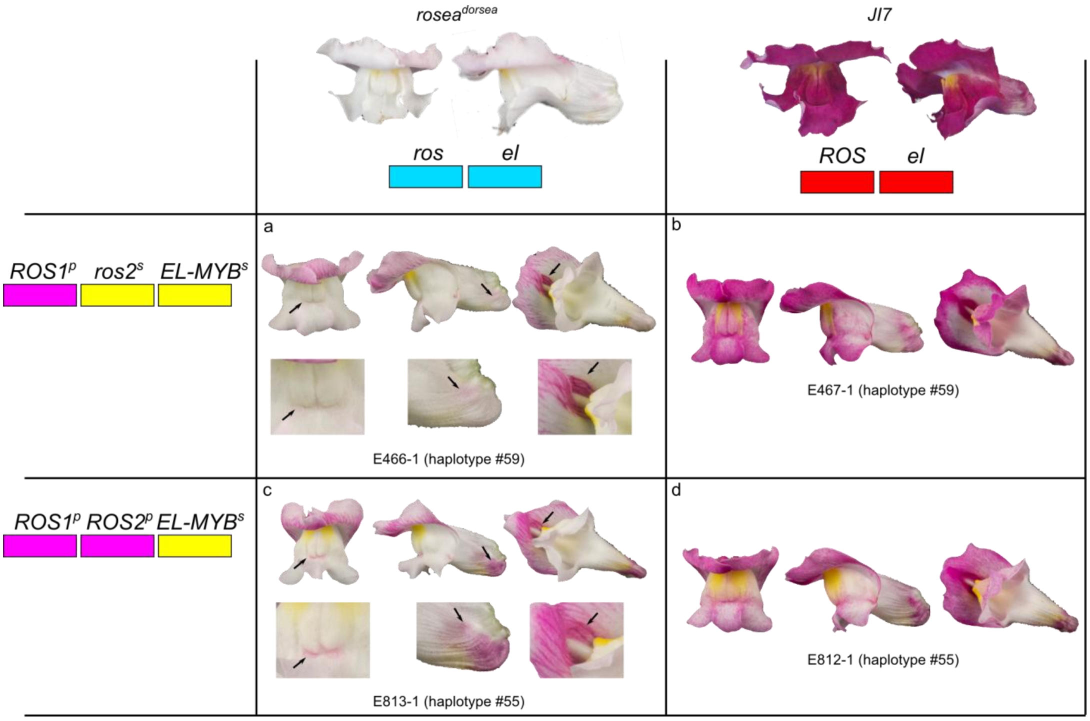


Figure 6.14 – Influence of recombination points between *ROS1* and *ROS2* on pigment production.

Representative photographs are shown for each case, with haplotype numbers corresponding to the attached file "*hz_haplotypes.xlsx*". Arrows in panels a and c point to regions where Eluta flowers are typically pigmented and are emphasized below the photograph of the whole flower. The haplotypes for each gene are represented with coloured boxes to indicate the origin of the allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea*^{dorsea}.

This observation provides further evidence that elements downstream of *ROS1* (perhaps *ROS2* or enhancers of *ROS1* expression in its 3’ untranslated region), contribute to some of the pigmentation in these flowers, which agrees with the previously discussed genetic analysis of the rosy phenotype (section 4.2.2). The effect of *EL* in these recombinants seems to be unchanged, since the F1 test-cross to *JI7* results in a similar phenotype, despite their different *ROS1-ROS2* combination (Figure 6.14 b and d).

Further evidence of the involvement of elements linked to *ROS1* in producing small amounts of pigment comes from the analysis of 5 haplotypes *ros1*^{s} *ros2*^{s} *el-myb*^{p}. The test-cross progeny with this haplotype is expected to produce non-magenta flowers when crossed to *rosea*^{dorsea}, as some examples show (Figure 6.15 c-d). However, 2 out of the 5 *ros1*^{s} *ros2*^{s} *el-myb*^{p} haplotypes produced some pigment in the F1 with *rosea*^{dorsea} (Figure 6.15 a-b). The pigmentation pattern is similar to the rosy phenotype (albeit lighter), suggesting that the same genetic elements could be involved in producing some pigment in these individuals. However, unlike the prediction that this should be associated with a recombination between *ROS1* and *ROS2*, these recombinants do not differ in the marker alleles of these genes. Since no further markers were used to characterize these recombinant haplotypes, these elements remain to be fine-mapped. However, this result suggests such elements may in fact be located downstream of *ROS2* (i.e. past the marker used).

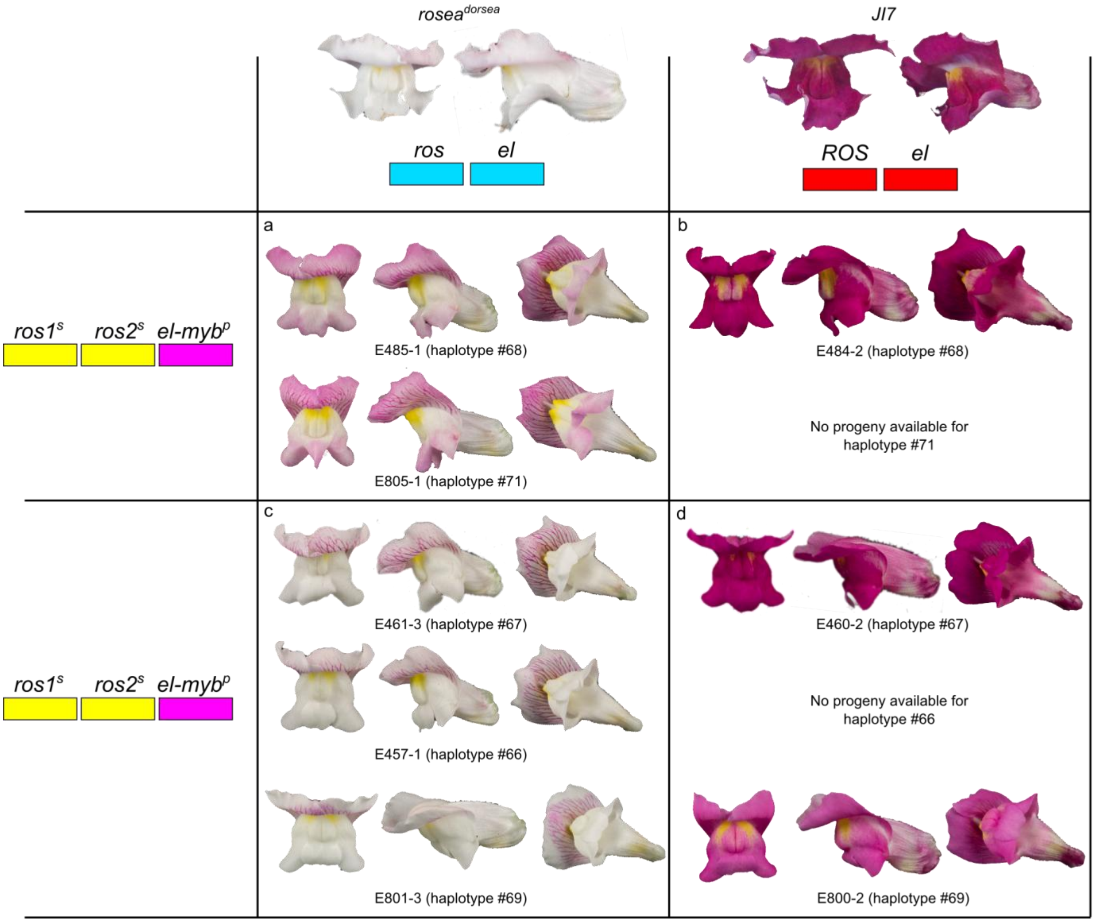


Figure 6.15 – Unexpected production of pigment by *ros1^{s} ros2^{s} el-myb^{p}* haplotypes. Representative photographs are shown for each case, with haplotype numbers corresponding to the attached file "*hz_haplotypes.xlsx*". The haplotypes for each gene are represented with coloured boxes to indicate the origin of the allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea^{dorsea}*.

Finally, one recombinant haplotype was obtained with what seems to be a product of a double-recombination event: this haplotype carries *pseudomajus* alleles for *ROS1* and *EL-MYB*, but a *striatum* allele for *ROS2* (*ROS1^{p} ros2^{s} el-myb^{p}*). In the F1 test-cross to *JI7*, individuals have magenta flowers, indistinguishable from crosses involving a non-recombinant *pseudomajus* haplotype (Figure 6.16 compare b and d). Unfortunately, only the test-cross progeny to *JI7* was obtained for this haplotype, thus its influence on the phenotype could not be assessed in the recessive background of *rosea*^{dorsea}. This would allow the individual contribution of *ROS1*^{p} to the pigment of flowers to be distinguished.

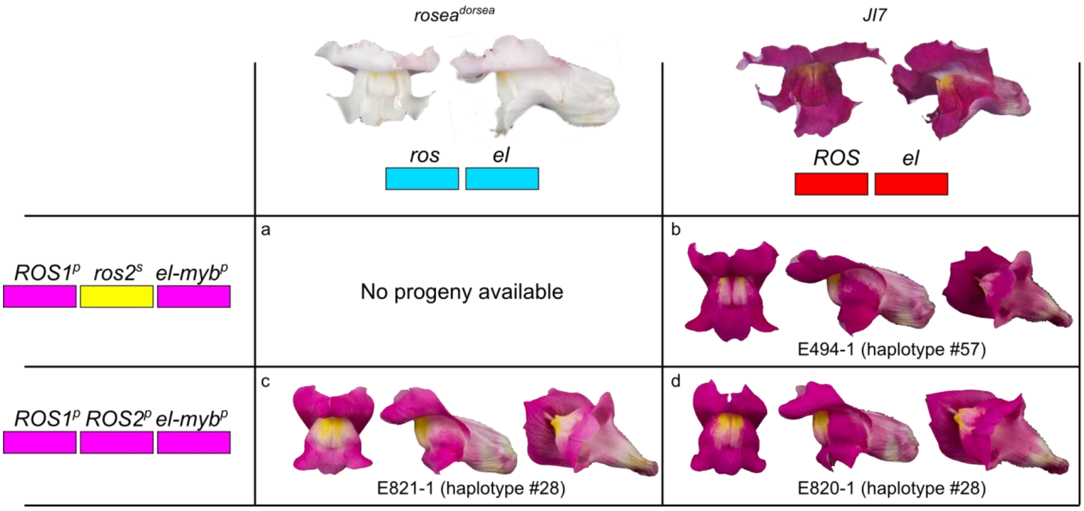


Figure 6.16 – Phenotype conferred by a *ROS1*^{p} *ros2*^{s} *el-myb*^{p} double recombinant. Representative photographs are shown for each case, with haplotype numbers corresponding to the attached file "*hz_haplotypes.xlsx*". The haplotypes for each gene are represented with coloured boxes to indicate the origin of the allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea*^{dorsea}.

Taken together, the analysis of these haplotypes suggests that one or more elements linked to *ROS1-ROS2* may contribute to the control of anthocyanin pigmentation in the flowers. In all cases described in this section, the influence of the putative elements linked to *ROS1-ROS2* is not visible in crosses to *JI7*, suggesting that they are recessive to the *ROS*^{7} allele but dominant to the *ros*^{dor} allele.

The influence of these elements on the phenotype of the flowers may explain some of the phenotypes observed in the hybrid zone population. Namely, 7% of plants with the genotype *ros1*^{s} *EL-MYB*^{S}/*ros1*^{s} *el-myb*^{P}, which are expected to be non-magenta, actually appear to be Eluta (Figure 6.6B). One hypothesis is that those plants are able to produce some pigment in the flowers because they carry *pseudomajus* elements downstream of *ROS1*.

6.2.5.4 Functional polymorphism in EL may be located upstream of EL-MYB

Typical *striatum* haplotypes (*ros EL*) should result in the Eluta phenotype when heterozygous with *JI7* (Figure 6.17d). This was the case for 16 out of 18 *ros1^s ros2^s EL-MYB^s* parental haplotypes analysed (Figure 6.9h). However, two exceptional *striatum* haplotypes produced magenta F1 progeny when crossed to *JI7* (Figure 6.17b), suggesting, in fact, that they carry a recessive allele of *EL*. This discrepancy between the molecular and functional genotypes suggests that *EL* could be located beyond the marker used in the *EL-MYB* gene. This marker is located in the promoter region of *EL-MYB* (Figure 6.10), thus the most likely explanation for this result is that the causative *EL* polymorphism is further upstream in the promoter region of this gene (notice that due to the orientation of this gene, this is equivalent to being further downstream of *ROS1*).

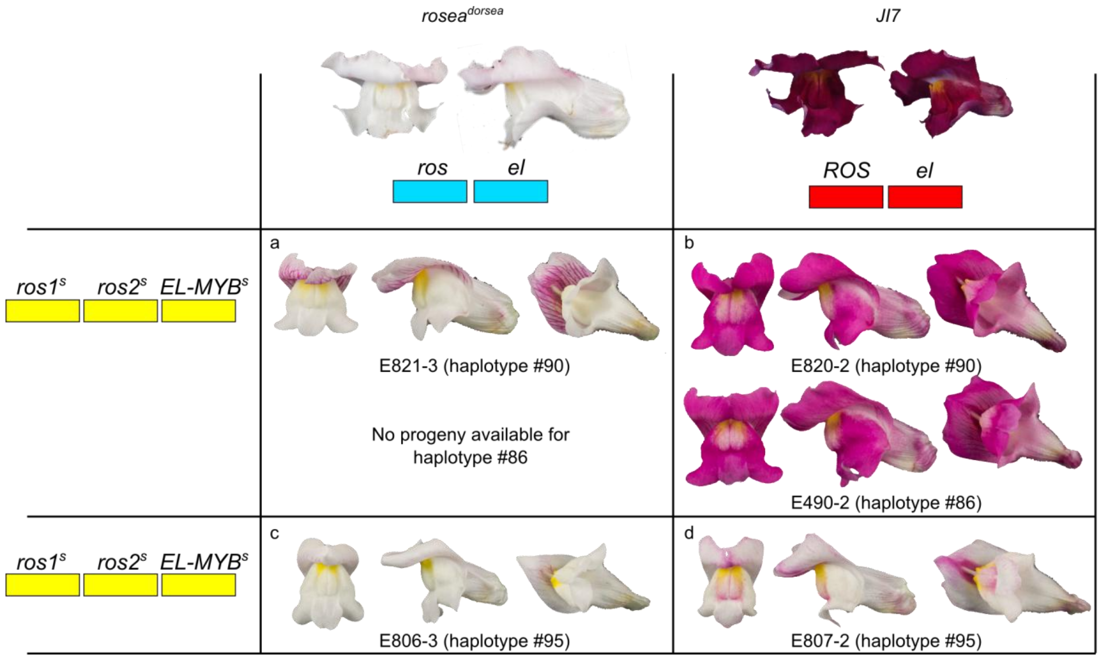


Figure 6.17 – *striatum* molecular haplotypes that do not carry *EL*.
Two haplotypes with *striatum* alleles in the molecular markers produce magenta progeny when crossed to *JI7* (haplotypes #90 and #86). This is distinct from all other similar haplotypes (exemplified by haplotype #95), which produce Eluta flowers in this test-cross. Representative photographs are shown for each case, with haplotype numbers corresponding to the attached file "*hz_haplotypes.xlsx*". The haplotypes for each gene are represented with coloured boxes to indicate the origin of the allele. Colours are: magenta – *pseudomajus*; yellow – *striatum*; red – *JI7*; blue – *rosea*^{dorsea}.

This observation may explain some of the genotype-phenotype discrepancies observed in the hybrid zone (Figure 6.6). For example, 2% of heterozygous individuals with the two parental haplotypes (*ROS1*^{p} *el-myb*^{p} / *ros1*^{s} *EL-MYB*^{S}) are magenta instead of the expected Eluta. This result could be attributed to cases where a recombination occurred between the marker in *EL-MYB* and the actual functional site of *EL*. Indeed, one of the haplotypes analysed in the test-crosses (haplotype #90; Figure 6.17a-b) derives from a hybrid zone individual carrying both parental haplotypes (*ROS1*^{p} *el-myb*^{p}/*ros1*^{s} *EL-MYB*^{S}) but not showing an Eluta phenotype (Figure 6.18).

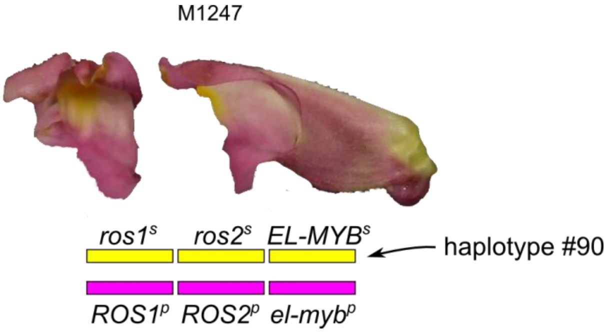


Figure 6.18 – Hybrid zone individual carrying an *A. m. striatum* molecular haplotype with a recessive allele of *EL*.

This individual is the hybrid zone parent that gave origin to haplotype #90 that, in terms of function, carries a recessive *el* allele (see Figure 6.17b). Although the flower of the individual is partially damaged, it is clear that the phenotype is magenta and not Eluta as would be expected from its genotype.

6.2.6 Interaction between *EL* and vein pigmentation

Up to this point, the magenta pigmentation that specifically overlaps the veins of the petals has not been considered for the floral phenotype (Figure 6.19A). However, it is important to consider it, since it is known to influence pollinator behaviour in *A. majus* (Shang *et al.* 2011). The phenotype resultant from this vein-specific pigmentation (onwards referred to as venation; Figure 6.19A) confers a striped pattern of pigmentation in the dorsal petals of *Antirrhinum*. The pigmentation of veins occurs only in the inner surface of the dorsal petals' lobes, being mostly absent from other regions of the flower (some venation can be observed inside the tube of the flower, but this was not considered in this study). This pattern is clearly visible in a non-magenta background (*ros/ros*), but becomes cryptic in a magenta background (*ROS/ROS*).

Venation is controlled independently of *ROS-EL* by an unlinked gene named *VENOSA* (*VEN*), which also encodes a *MYB*-like transcription factor (Schwinn *et al.* 2006; Shang *et al.* 2011). The *JI7* and *rosea*^{dorsea} lines used for the genetic analysis in test-crosses have a recessive non-functional allele of *VEN* and therefore the veins are not pigmented in these plants (non-venation phenotype). However, in the test-crosses using individuals from the hybrid zone heterozygous with *rosea*^{dorsea}, venation is clearly visible in some flowers, suggesting that dominant functional alleles of *VEN* occur in the hybrid zone.

VEN has not been studied intensively in the hybrid zone, and in fact the pigmentation overlapping the veins is not considered in the scoring system used for flower colour phenotypes. However, there is evidence that both functional and non-functional alleles of this locus segregate in the hybrid zone (Figure 6.19B). In the progeny of a cross between an *A. m. striatum* and a *rosea*^{dorsea} (recessive *ven/ven*), all flowers were non-magenta, but segregated in a 1:1 ratio for the venation and non-venation phenotypes. This suggests that the parental *A. m. striatum* individual used in this cross was heterozygous *VEN/ven*.

Although a systematic scoring of the venation phenotype has not been considered in the hybrid zone, *A. m. striatum* individuals observed in the field typically have their veins restricted to the junction of the two dorsal petals (see, as an example, individual Y139-3 in Figure 6.19B). However, in other backgrounds, the pigmentation of the veins is often spread outside of that region (for example, Figure 6.19A).

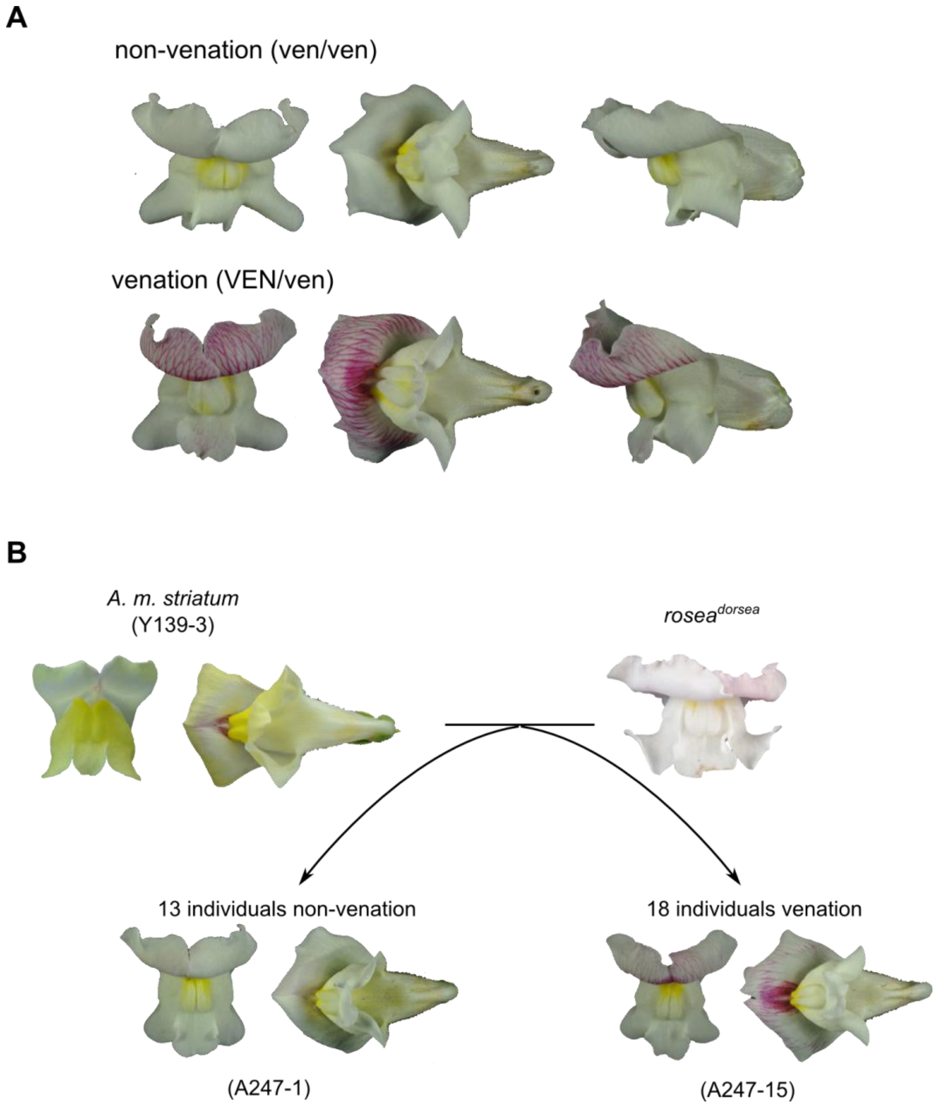


Figure 6.19 – Vein-specific pigmentation in A. majus.

A) Example of flowers with and without pigmentation overlapping the veins. The individual on the top is homozygous recessive *ven/ven* and the individual on the bottom heterozygous *VEN*/*ven*. Pigmentation occurs in the epidermis overlapping the veins in the inner surface of the dorsal petals only. These individuals are not from the hybrid zone.

B) Segregation of vein pigmentation in a cross between an *A. m. striatum* and *rosea*^{dorsea}. The *A. m. striatum* individual used in this cross (Y139-3) was obtained by crossing two individuals grown from wild seed, caught ~13km away from the hybrid zone centre. Therefore, it should represent a typical *A. m. striatum* from parapatric populations of this subspecies. The F1 of the cross to *rosea*^{dorsea} produced progeny with and without pigmented veins in a 1:1 ratio (χ^2 = 0.81, p = 0.37), suggesting the *A. m. striatum* parent was heterozygous *VEN/ven*. Photographs of representative phenotypes in the F1 are shown, with the ID of the individual in the photograph given for reference.

Both *ROS1* and *EL-MYB* are *R2R3 MYB* transcription factors and, because *VEN* is also an *R2R3 MYB* gene, it could be postulated that *EL* genetically interacts with *VEN* in a similar way that it does with *ROS*. Therefore, the following hypothesis can be postulated: that a dominant allele of *EL* not only restricts the magenta pigment conferred by *ROS* but also that conferred by *VEN*.

To investigate the possibility of an interaction between *EL* and the venation phenotype, I analysed the venation patterns in the test-cross progeny of the haplotypes sampled from the hybrid zone. Because venation is only clearly visible in a non-magenta background, I only considered haplotypes recessive for *ros* (*ros EL* or *ros el*) heterozygous with *rosea*^{dorsea}. First, I developed a scoring system for the venation pattern, based on the variation seen across multiple test-cross progeny (Figure 6.20). Generally, veins may be visibly restricted to the region where the two dorsal petals meet (1 ≤ score < 3) or evenly spread across those petals (3 ≤ score ≤ 4). In the cases where venation is restricted, there can be some spread outside of the dorsal petals' junction, but the venation does not reach the lower part of the petals (scores 1 to 2.5 in Figure 6.20). Conversely, when venation is not visibly restricted to the dorsal petals' junction, the veins are pigmented from the lower part of the dorsal petals, although sometimes not reaching all the way to the upper border of the petals (scores 3 to 4 in Figure 6.20).

If *EL* interacts with the venation pattern, then *ros1*^{s} *EL-MYB*^{s}; *VEN* crosses to *rosea*^{dorsea} should result in vein restriction, whereas crosses with *ros1*^{s} *el-myb*^{p}; *VEN* should result in evenly spread veins. There were only 5 *ros1*^{s} *el-myb*^{p} haplotypes crossed to *rosea*^{dorsea}, but all those resulted in non-restricted veins (score ≥ 3; see venation in Figure 6.15 a and c). There were 16 <u>*ros1*^{S} *EL-MYB*^{S}</u> haplotypes crossed to *rosea*^{dorsea}, although only 14 cases could be scored for veins (possibly due to the presence of recessive *ven* alleles in some of the individuals used in the test-crosses, and so did not produce any venation). 12 of these 14 individuals had restricted veins (score < 3), which largely supports that the dominant allele of *EL* restricts venation to the region where the dorsal petals meet.

Score	Vein restriction	Vein spread
1	veins visibly restricted where the two dorsal petals meet	no visible spread onto the sides
2 - 2.5		some spread onto the sides, but not reaching lower part of the dorsal petals
3 - 3.5	veins not visibly restricted, but evenly spread across the lower part of the dorsal petals	veins do not reach the top of the dorsal petals
4		veins reach the top of the dorsal petals

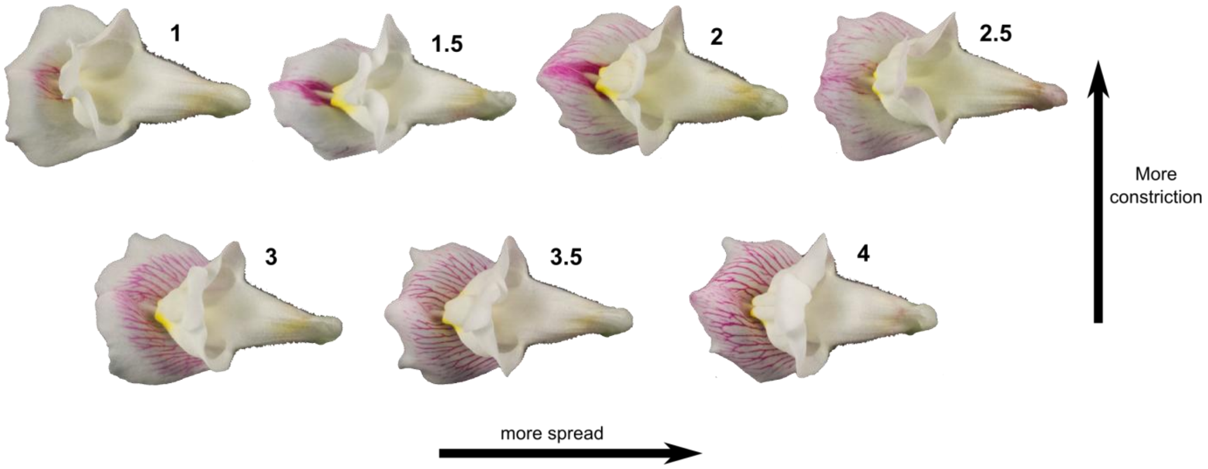


Figure 6.20 – Scoring system for vein pigmentation.
The table describes each scoring category in terms of the extent of vein restriction in the region where the dorsal petals meet and in terms of vein spread outside of that region. A bottom-up view of the flowers is shown, to emphasise the vein pigmentation in the dorsal petals.

There were two exceptions where *ros1*^{s} *EL-MYB*^{s} haplotypes resulted in spread venation in the test-cross: haplotype #73 with vein score 3.5 and haplotype #90 with vein score 4. Incidently, haplotype #90 was shown to actually carry a functionally recessive allele of *el* (Figure 6.17a-b; notice the spread veins in the cross to *rosea*^{dorsea}), again supporting the idea that *EL* restricts the venation pattern. Unfortunately, the other *ros1*^{s} *EL-MYB*^{s} haplotype with spread venation (haplotype #73, not shown) was not crossed to *JI7*, therefore its functional *EL* genotype could not be confirmed, but the prediction is that it may actually carry a recessive allele of *el*.

6.3 Discussion

6.3.1 Hybrid zones as a natural genetic resource for fine-mapping functional polymorphisms

In this work, I took advantage of naturally occurring recombination between *A. m. pseudomajus* and *A. m. striatum* to provide new genetic material for fine-mapping functional polymorphisms in the *ROS-EL* region. Individuals were first identified as recombinants by using molecular markers in the *ROS1* and *EL-MYB* genes, and then their progeny grown and crossed to *A. majus* lines of known genotype (Figure 6.8). In this way, the molecularly-identified recombinants from the hybrid zone could be confirmed as being functional recombinants between *ROS* and *EL*, providing genetic proof that functional *ros el* and *ROS EL* recombinants segregate in the hybrid zone.

This strategy also provided unique material to further dissect the genetic architecture of this complex cluster of loci. Firstly, it confirmed that the markers in *ROS1* and *EL-MYB* are tightly linked to the functional *ROS* and *EL* loci, respectively. However, this association is not perfect. For example, a magenta individual with heterozygous *ROS1*^{p} *el-myb*^{p}/*ros1*^{s} *EL-myb*^{s} genotype in the hybrid zone (Figure 6.18) revealed that the functional polymorphism in *EL* is likely upstream of the *EL-MYB* coding sequence, possibly in a regulatory element of this gene. Extending from the mapping experiments in chapter 4, there is further evidence that *ROS2* (or elements downstream of it) contribute to making small amounts of magenta pigment in the flowers (Figure 6.14). These cases demonstrate that by genetically tracking molecular recombinants between *ROS1* and *EL-MYB*, or cases where the molecular genotype does not match expectations about the phenotype in the hybrid zone, we can increase the fine-mapping resolution of functional loci involved in flower colour regulation.

The test-crosses with individuals from the hybrid zone also provided evidence that a putative dominant repressor of magenta pigment, unlinked to *ROS-EL*, segregates in the hybrid zone (Figure 6.12). This can be an important feature to consider in the hybrid zone, since the same *ROS-EL* haplotype may produce different phenotypic outcomes depending on the background it occurs. Indeed, the occurrence of non-magenta phenotypes in individuals expected to be Eluta (4% of *ROS1*^{p} *el-myb*^{p}/*ros1*^{s} *EL-myb*^{s} and 29% of *ros1*^{s} *EL-myb*^{s}/*ROS1*^{p} *EL-MYB*^{s}; Figure 6.6B) may be partially due to this putative repressor of anthocyanin in the background. The genetic material from this work should allow the mapping of this putative repressor in *Antirrhinum* and explore how its alleles segregate in the hybrid zone. Repressors of anthocyanin pigmentation have been cloned in other species, such as *ROSE INTENSITY1* in *Mimulus* (Yuan *et al.* 2013), *INTENSIFIER1* in maize (Burr *et al.* 1996), and *AtMYBL2* in *Arabidopsis* (Matsui *et al.* 2008). These can be also used as candidate genes for mapping.

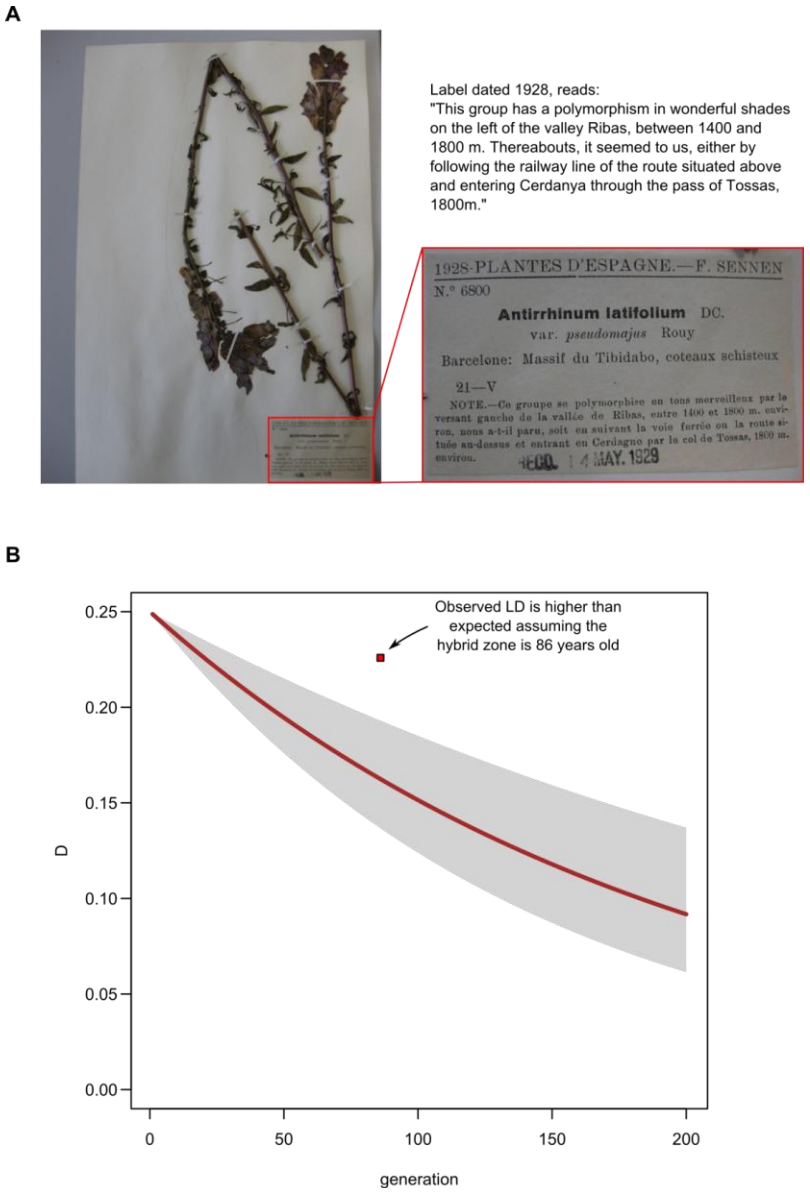
Natural populations are frequently used to map loci related to particular traits, in both animal and plant systems (Mackay & Powell 2007; Donnelly 2008; Myles *et al.* 2009). Usually, this aims at finding associations between particular genetic markers and the trait (or traits) of interest, a method known as association or linkage disequilibrium mapping. The main outcome of such experiments is the identification of previously unknown loci, which are then functionally characterized using various methods (e.g. gene expression analysis, functional knock-outs, over-expression, etc.). Even though these methods take advantage of naturally occurring variation to track down the genetic basis for trait differences between individuals in a population, it differs from the strategy used here for characterizing *ROS-EL* haplotypes. Rather than aiming at identifying unknown loci, prior knowledge about *ROS* and *EL* was used to obtain an extensive collection of haplotypes used for fine-mapping linked interacting loci responsible for a natural polymorphism.

Even though the mapping of *ROS-EL* can be achieved using traditional pedigree-based mapping experiments (chapter 4), using natural recombinants has the advantage that the haplotypes analysed derive from the actual population under study, segregating in the species genetic background. Unlike some of the glasshouse experiments, these recombinants involve naturally segregating alleles and haplotypes that are recombining due to the dynamics of gene flow between *A. m. pseudomajus* and *A. m. striatum* in the hybrid zone. Therefore, there is no extrapolation needed between the glasshouse experiments and the observations in the field. To my knowledge, this is a unique case where a natural population was used as a source of genetic material for fine-mapping linked interacting loci. This material should prove an invaluable resource for future identification of functional regions that characterize the *A. m. pseudomajus* and *A. m. striatum ROS-EL* haplotypes.

6.3.2 Ongoing recombination between *ROS* and *EL* in natural populations

The survey of a large sample of individuals from the hybrid zone population revealed that a significant number of recombinant $\underline{ROS1^{p}\ EL\text{-}MYB^{s}}$ and $\underline{ros1^{s}\ el\text{-}myb^{p}}$ haplotypes are segregating in the population. The genetic analysis of some of these haplotypes revealed that, save some exceptions, the markers in *ROS1* and *EL-MYB* are good indicators for the genotype in the functional *ROS* and *EL* loci, respectively (Figure 6.9). Therefore, I will consider some general conclusions by focusing on the functional (rather than molecular) *ROS-EL* haplotypes.

The occurrence of recombinant *ROS-EL* haplotypes in the hybrid zone is not wholly unexpected if gene flow has been ongoing for a sufficient number of generations (Hedrick 2011; Wang *et al.* 2011). In fact, under random mating and no population structure, linkage disequilibrium (LD) between adjacent loci is expected to decay over multiple generations, as recombinant haplotypes accumulate in a population. Based on this neutral model for LD decay, the observed LD between *ROS* and *EL* in the central region of the hybrid zone ($D = 0.22$) would predict that recombinant haplotypes would reach the observed frequency of ~5% in only 25 years (= generations). However, there is an alternative source of evidence suggesting that the hybrid zone may in fact be older than this estimate. Christophe Thébaud (Univ. Toulouse) found a herbarium specimen of *A. majus* collected in 1928 by the French botanist Sennen, in which the label reads: “This group has a polymorphism in wonderful shades on the left of the valley Ribas, between 1400 and 1800m” (Figure 6.21A). This description not only implies that Sennen must be referring to the segregation of flower colour phenotypes in a hybrid zone, but the locality he refers to coincides to where the *pseudomajus* x *striatum* hybrid zone is found nowadays. This increases the estimated age of the hybrid zone to being at least 86 years old. This means that the observed LD between *ROS* and *EL* is in fact higher than that estimated from the neutral model considered (Figure 6.21B). This might be due to selection acting against recombinant haplotypes, but can also have a non-selective explanation related to the structure of the population.



Label dated 1928, reads:
"This group has a polymorphism in wonderful shades on the left of the valley Ribas, between 1400 and 1800 m. Thereabouts, it seemed to us, either by following the railway line of the route situated above and entering Cerdanya through the pass of Tossas, 1800m."

1928-PLANTES D'ESPAGNE.—F. SENNEN

N.° 6800

Antirrhinum latifolium DC.
var. *pseudomajus* Rouy

Barcelone: Massif du Tibidabo, coteaux schisteux

21—V

NOTE.—Ce groupe se polymorphise en tons merveilleux par le versant gauche de la vallée de Ribas, entre 1400 et 1800 m. environ, nous a-t-il paru, soit en suivant la voie ferrée ou la route située au-dessus et entrant en Cerdagne par le col de Tossas, 1800 m. environ.

RECD. 14 MAY. 1929

Observed LD is higher than expected assuming the hybrid zone is 86 years old

Figure 6.21 – Estimates for the age of the hybrid zone.

A) Photograph of a herbarium specimen of *A. m. pseudomajus* (formerly *A. latifolium* var. *pseudomajus*) found in the Natural History Museum in London. The label description suggests that the collector found the hybrid zone considered in this study, dating it back to at least 86 years. Photograph courtesy of Sandra Knapp (NHM, London).

B) Observed and expected linkage disequilibrium between *ROS* and *EL*, ignoring spatial structure and selection. The line and shaded area represent the theoretical decay of $\boldsymbol{D}$ over the generations, assuming random mating and no selection against recombinant haplotypes. From three different mapping experiments detailed in chapter 4, *ROS* and *EL* were estimated to be 0.3cM, 0.5cM and 0.7cM apart. Based on this, the grey shaded area defines the decay of $\boldsymbol{D}$ for recombination rates between 0.003-0.007 and the line for a recombination rate of 0.005. The point plots the observed value of $\boldsymbol{D}$ between *ROS1* and *EL-MYB* in the hybrid zone, assuming a minimum age of 86 years for the hybrid zone and one generation a year.

The model of LD decay used here, ignored the spatial structure of the hybrid zone, and therefore it assumes that recombinants can form anywhere across the geographic transect (although I restricted my calculation of LD within 500m of the centre). However, the haplotype frequencies are not homogeneous across the hybrid zone (Figure 6.5B), and therefore recombinants form mostly in the central region, where heterozygotes occur at a highest frequency. Although recombinant haplotypes can be found in the flanking region of the hybrid zone, this is likely because they spread out from the centre over time. Furthermore, migration of individuals from the flanks into the centre of the hybrid zone might also inflate the levels of LD, because they will preferentially carry *A. m. pseudomajus* or *A. m. striatum* haplotypes. Therefore, more realistic models that consider the particular structure of hybrid zones should be used in the future, to infer if LD between *ROS* and *EL* is higher than would be expected without selection and considering that the present hybrid zone might be ~90 years old (based on herbarium evidence).

Another component of the spatial structure of this population is the small scale distribution pattern of individuals in the field. In the studied *Antirrhinum* population, individuals seem to have a patchy, rather than homogeneous, geographic distribution, with individuals growing mainly on the edges of human-made roads. In fact, *Antirrhinum* thrive on disturbed habitats as, for example, the exposed rocky substrate left after the construction of a road. Conversely, shady areas, with dense tree vegetation, are often bare of any *Antirrhinum* plants. This can lead, for example, to local founder effects, resulting in patches composed of low genetic diversity. It can also constitute a barrier to pollen exchange (mediated by bee species), such that flowers receive more pollen from plants in the same patch than from plants from neighbouring patches (Turner *et al.* 1982; Rasmussen & Brødsgaard 1992). I have not investigated in detail if the geographic distribution of individuals influences allele frequencies at a local level. However, this analysis can be done in the future by analysing allele frequencies within areas of varying sizes to see if they deviate from the expectation of random mating (Hardy-Weinberg equilibrium). For example, patches of individuals may contain a low diversity of haplotypes with mostly homozygous individuals, leading to an increase of LD between *ROS* and *EL*.

Finally, it would be important to investigate how long *A. majus* seeds can remain dormant in the soil. Dormant seeds that persist in the soil for several years (forming soil seed banks), constitute an important source of genetic variation, which does not necessarily reflect current, but past demographic states of the population (Mandák *et al.* 2006). Furthermore, the genetic diversity found in seed banks might itself be the product of selection (e.g. seed viability might depend on its genotype). Thus, allele and haplotype frequencies in the hybrid zone might be influenced by the genetic diversity present in the seed bank.

Although explanations that do not invoke any selection have to be considered, the observation of a strong signal of high *Fst* in both *ROS* and *EL*, suggests that selection acts to maintain the *A. m. pseudomajus* and *A. m. striatum ROS-EL* haplotypes, despite gene flow and recombination in the hybrid zone. Therefore, hypotheses about the mechanisms and forms of selection have to be considered.

6.3.3 Hypotheses on selection against *ROS-EL* recombinants

One likely way in which flower colour may be under selection is related to the pollination biology of *Antirrhinum*. Pollination in this genus is carried out by large bee species, such as bumblebees (*Bombus* spp.) and carpenter bees (*Xylocopa spp.*) (Tastard *et al.* 2008; Vargas *et al.* 2010). These large insects are able to open the closed flowers of *Antirrhinum* to access nectar stored in the base of the tube of the flowers. While doing so, they rub their body against the pollen-containing stamens, and pollen is thus transmitted from plant to plant during the pollinator's foraging bout. Therefore, pollinator behaviour is one of the most likely agents of selection on flower colour in this system. It should be noted that bees can also see in the UV spectrum of light (Kevan *et al.* 1996; Chittka & Raine 2006) and therefore, the visually scored phenotypes from the hybrid zone may not capture all of the variation perceived by the pollinators. However, *A. m. pseudomajus* and *A. m. striatum* petals have been shown to have low reflectance levels in the UV and therefore the variation visible in the human visual spectrum is likely to reflect the main variation seen by the pollinators (Tastard *et al.* 2008).

One hypothesis of how pollinators could play a role in maintaining the hybrid zone is to consider that different species of pollinators preferentially pollinate one of the subspecies over the other (as well as any hybrid phenotypes). However, the same bee species are found pollinating both *A. m. pseudomajus* and *A. m. striatum* (Tastard *et al.* 2008) and therefore divergent pollinator preference to either subspecies is unlikely to play a role in the maintenance of this hybrid zone.

Another hypothesis is that pollinators learn to associate certain phenotypes with a reward (e.g. nectar or pollen) and keep visiting similar flowers after that learning period. This behaviour, known as flower constancy, has been observed in several pollinating insects, including bumblebees (Chittka & Thomson 1997; Gegear & Laverty 2005). Considering this behaviour, one hypothesis is that the fitness conferred by a particular floral phenotype increases with the frequency of similar phenotypes in neighbouring plants (i.e. there is positive frequency-dependent selection against rarer phenotypes). For example, a pollinator might be more likely to learn a phenotype-reward association with the most common phenotype found in a particular place (Smithson & Macnair 1996). This could lead to strong selection against rare phenotypes in the flanking regions of the flower colour cline where either the magenta (in the East) or the non-magenta (in the West) phenotypes are most common. Such frequency-dependent selection could explain the maintenance of the *ROS el* and *ros EL* haplotypes, while counteracting the spread of recombinant haplotypes. To explain this idea, it is useful to consider how each recombinant haplotype influences the phenotype in either subspecies background.

If a *ROS EL* haplotype was segregating in an *A. m. pseudomajus* background, it would produce Eluta flowers (see *ROS el* / *ROS EL* in Figure 6.6B). This phenotype would be rare in comparison with the prevalent magenta phenotype of *A. m. pseudomajus*, because recombinant haplotypes form in the centre of the hybrid zone and slowly spread to the flanks. Therefore, under the positive frequency-dependent selection hypothesis, individuals carrying such haplotype would incur a fitness cost in the *A. m. pseudomajus* background. This fits with the observation that *ROS EL* haplotypes are rarely found in the Eastern side of the flower colour cline (only 4 haplotypes were found between 500-850m East-West of the hybrid zone centre, and none further from that distance; Figure 6.5B). However, the flanks of the hybrid zone were not sampled as intensively as the central region, so this result may be related to a lack of power to detect such introgression.

On the other hand, a *ros el* haplotype in the *A. m. pseudomajus* background would produce magenta flowers (see *ROS el* / *ros el* in Figure 6.6B). Being phenotypically indistinguishable from the *A. m. pseudomajus* phenotype, individuals carrying this haplotype might in fact not be selected against. However, if the *ros el* haplotype was to increase in frequency locally, the probability of forming homozygotes *ros el*/*ros el* would also increase. Such homozygotes would result in non-magenta flowers (*ros el*/*ros el*), and therefore would be negatively selected against compared to the common magenta phenotype. Therefore, a *ros el* haplotype could exist in the *A. m. pseudomajus* background, as long as its frequency remained relatively low (such that it occurs mostly in heterozygous form). Despite this, there is no evidence that this haplotype occurs at significant frequencies in the magenta flank, since it was very rarely observed in the eastern flanking region of the hybrid zone (no such recombinants were found further than 618m East of the hybrid zone centre; Figure 6.5B). However, as mentioned above, this may be related to a lack of power to detect rare haplotypes in this area.

In the *A. m. striatum* population, a *ROS EL* recombinant haplotype will produce mostly Eluta flowers (see *ros EL*/*ROS EL* in Figure 6.6B). As above, this would reduce the fitness of individuals carrying this haplotype, since the prevalent phenotype in the *A. m. striatum* population is non-magenta. However, unlike what occurs in the *A. m. pseudomajus* background, these individuals are homozygous *EL*/*EL*. This is significant because of the semi-dominance of this locus, which results in a stronger reduction of the magenta pigment in the flowers compared to when they are *EL*/*el* heterozygous. Indeed, a significant fraction (29%) of *ros EL*/*ROS EL* individuals in the hybrid zone have a non-magenta phenotype (Figure 6.6B). This effect may be exacerbated since there is evidence that *ROS* may not be fully dominant (i.e., one copy of the dominant allele promotes the production of less pigment than two copies; Figure 6.11). Moreover, the presence of modifiers of pigment unlinked to *ROS*-*EL* may further reduce the pigmentation promoted by *ROS* (Figure 6.12), if they occur in the *A. m. striatum* background. Taken together, these observations suggest that selection on the *ROS EL* haplotype might not be as severe in the *A. m. striatum* background, since individuals may actually resemble the prevalent non-magenta phenotype. Supporting this observation, there is some evidence for introgression of this haplotype into the Western side of the hybrid zone (21 of such haplotypes are found beyond 500m East-West of the hybrid zone centre, and going as far as 6km).

Finally, the *ros el* recombinant haplotype in the *A. m. striatum* background will produce non-magenta flowers (see *ros EL*/*ros el* in Figure 6.6B). Similarly to what happens in the *A. m. pseudomajus* background, this would make this haplotype relatively neutral in this context, since its phenotype might resemble that of the non-magenta *A. m. striatum* individuals. However, these flowers may have an altered venation pattern, as suggested by the interaction between *EL* and the spread of vein-specific pigmentation in the dorsal petals of the flowers (Figure

6.20). In addition, due to the semi-dominance of *EL*, homozygotes *ros el*/*ros el* may have a more severe spreading of venation. If these differences in venation pattern are distinguished by pollinators, then individuals with spread venation might be selected against in the *A. m. striatum* background. The interaction between *EL* and venation will have to be confirmed more carefully using crosses in the subspecies genetic background (as opposed to a *rosea*^{dorsea} background). These experiments are underway and should provide the confirmation that *EL* is likely to interact with this component of the floral phenotype.

In summary, under the positive frequency-dependent hypothesis of selection, both *ros el* and *ROS EL* haplotypes would be selected against. This may particularly apply to the flanks of the hybrid zone, where the occurrence of a prevalent phenotype makes it harder to increase the frequency of phenotypes conferred by recombinant *ROS-EL* haplotypes.

There is a second likely form of selection, also mediated by pollinator behaviour, which could explain the maintenance of the subspecies *ROS-EL* haplotypes. It could be that pollinators have a biased preference for certain phenotypes, over others. This hypothesis is frequency independent; that is, the fitness of an individual does not depend on the phenotype of surrounding individuals, but rather on the preferential behaviour of the pollinator. For example, regardless of prior learning, pollinators might prefer magenta (*A. m. pseudomajus*) and non-magenta with restricted venation (*A. m. striatum*) flowers over Eluta (conferred by *ROS EL*) and non-magenta with spread venation (conferred by *ros el*) flowers.

Both frequency-dependent and frequency-independent forms of selection could account for the maintenance of the allelic combinations of *ROS-EL* seen in the subspecies. These forms of selection are not mutually exclusive and, in fact, there is empirical evidence supporting that both may operate in *Antirrhinum*. For example, experiments using artificial arrays of *A. majus* with different phenotypes show that bumblebees do not randomly pollinate flowers, but rather mostly visit flowers of the same phenotype within each foraging bout (i.e., they show flower constancy behaviour; Niovi Jones and Reithel 2001; Oyama et al. 2010). Moreover, preliminary experiments using artificial arrays composed of *A. m. striatum* and *A. m. pseudomajus* plants at different frequencies, suggest that the visitation of different phenotypes may be influenced by the density of each phenotype (Tom Ellis, pers. comm.). These experiments offer support for the hypothesis that frequency-dependent selection (particularly in the flanks of the hybrid zone, where one type of phenotype is more common) might contribute to the maintenance of the *ROS el* and *ros EL* haplotypes.

On the other hand, pollinators have certain innate colour preferences (Lunau & Maier 1995; Lunau *et al.* 1996), which could support a frequency-independent form of selection if there is a biased attraction towards certain phenotypes. For example, field trials using arrays of *A. majus* plants of different phenotypes have shown that bumblebees visit magenta flowers significantly more often than white flowers (Shang *et al.* 2011). Moreover, *A. majus* flowers with venation and non-venation phenotypes can be distinguished by naive bumblebees (i.e., reared bumblebees that had no prior contact with any flowers) and are thought to serve as nectar guides by these pollinators (Shang *et al.* 2011; Whitney *et al.* 2013).

Finally, other species of plants that are visited by the same pollinators as *A. majus*, might also interfere with the fitness of individuals in the hybrid zone. On the one hand, there might be competition for pollination, that is, the quantity of visits to *A. majus* plants might be reduced if the pollinators visit other species. If pollinators shift between species, the quality of the visits might also decrease, affecting both male fitness - in the case where pollen is exported to a non-conspecific pistil - and female fitness - if the pollen received in the pistil is from another species (Mitchell *et al.* 2009). The opposite effect of competition, that is, facilitation of pollination, might also occur: if several species are similarly attractive to pollinators, this might increase the pollinator number in the community and, thus, increase the probability of visitation (although inter-specific pollen flow might counteract this advantage, unless the different species flower at different times). In this context, it would be desirable to characterize the species that share pollinators with *A. majus* in the hybrid zone and see if their distribution varies across the transect. Note that this ecological effect can affect the fitness both in the frequency-dependent and frequency-independent modes of selection described above.

In summary, the observation that *ROS* and *EL* are in high linkage disequilibrium in the hybrid zone and that a strong signal of high *Fst* is observed in both loci, suggests that selection acts to maintain the *A. m. pseudomajus* and *A. m. striatum ROS-EL* haplotypes, despite gene flow and recombination in the hybrid zone. The detailed genetic analysis of naturally derived haplotypes provides further evidence that multiple elements within this region contribute to the final phenotype of the flowers (e.g. *ROS1, ROS2* and *EL-MYB*). By considering the phenotypic consequences of recombination in *ROS-EL*, two forms of selection (frequency-dependent and frequency-independent) were proposed as ways in which recombinants could be selected against in the hybrid zone.

7 General discussion

7.1 Genomic divergence is shaped by a long history of recombination

I have demonstrated that the *ROS-EL* region is highly divergent (high *Fst*) between samples originating from the two flanks of a hybrid zone between *A. m. pseudomajus* and *A. m. striatum*. The divergence in these loci contrasts with the overall level of *Fst* across the genome, which is comparatively very low. This might be related to the homogenizing effect of gene flow between the two populations, or related to shared ancestral variation, due to a recent common ancestor. Those loci that are under selection (e.g. controlling flower colour) form sharp clines across hybrid zones between the two subspecies and are thus maintained differentiated between the populations. The observed heterogeneous pattern of *Fst* across the *ROS-EL* region fits the now prevalent view that genomes are permeable to gene flow between populations (which lowers *Fst* across the genome), but divergence can be maintained in a relatively few loci due to selection (Nosil *et al.* 2009; Strasburg *et al.* 2012). The selected loci form “islands” of high divergence punctuating a flat “sea” of (relatively) lower divergence (Turner *et al.* 2005; Nosil *et al.* 2009).

At a fine-scale, the *ROS-EL* region contains three prominent peaks of *Fst*, where most of the fixed polymorphisms in this region between *A. m. pseudomajus* and *A. m. striatum* samples are located (Figure 5.10A). These peaks tightly coincide with functional loci controlling different aspects of the magenta phenotype of the flowers in *Antirrhinum* (Figure 5.10B). The occurrence of fixed polymorphisms between samples from opposite sides of the hybrid zone is relatively rare and they mostly coincide with candidate genes within the fine-mapped intervals: *ROS1*, *ROS2* and *EL-MYB*. This result provides strong evidence that the three functionally identified loci are all under selection. To my knowledge, this is a rare case where tightly linked *Fst* peaks have been finely matched with functional loci responsible for a divergent trait. Although scans of genomic divergence are often noisy and should be interpreted with care (Beaumont 2005; Via 2012), this work demonstrates how they gain greatly from being complemented with genetic mapping experiments.

A similar case to the multi-peak divergence across the *ROS-EL* region in *Antirrhinum* is found in loci that control wing colour polymorphisms in the mimetic *Heliconius* butterflies. For example, the analysis of two races of *H. melpomene* with different phenotypes revealed that the region controlling the red patterning of the wings (*B/D* region) contains two peaks of divergence (*Fst*) within a region of approximately 200kb (~1cM) (Nadeau *et al.* 2012). One of these peaks coincides with the gene *optix*, involved in controlling this trait, whereas the other peak contains several uncharacterized polymorphisms, which are thought to consist of cis-regulatory elements of that gene (Reed *et al.* 2011; Martin *et al.* 2014; Pardo-Diaz & Jiggins 2014). However, in contrast with the present work on *Antirrhinum*, fine-mapping of the linked loci to tease apart their role in establishing the phenotypic differences between butterflies has thus far not been accomplished. This is possibly due to a limitation of rearing large numbers of butterflies to do genetic mapping, which is usually limited to a few hundred individuals (e.g. Jiggins and McMillan 1997; Joron et al. 2006; Baxter et al. 2008), rather than the several thousand that might be needed for mapping very tightly linked loci. Nonetheless, there are parallels between *A. majus* and *H. melpomene*, since both involve linked clusters of loci controlling traits that establish a reproductive barrier between distinct populations. Divergence between these loci is maintained, despite current and historical gene flow and recombination in hybrid zones that form between populations.

The fine-mapping experiments from this work revealed that the multiple, prominent divergence peaks in the *ROS-EL* region are unlikely due to noisy data, but rather pinpoint functionally important loci. How might such a rugged pattern of divergence occur? Likely it is due to the interplay between two opposing forces: selection and recombination. Selection maintains and/or increases divergence in a selected locus, but also in physically linked regions that "hitchhike" along with it (Feder & Nosil 2010; Flaxman *et al.* 2013). Therefore, a divergence peak is likely to contain a mixture of polymorphisms that are under selection and others that are neutral (they are in linkage disequilibrium in the population). However, recombination will start to break the associations between neutral and selected loci (linkage disequilibrium decays), resulting in narrower divergence peaks that more finely coincide with the selected locus. This recombination can occur between haplotypes segregating within a population or haplotypes from two different populations that hybridize. If several linked loci are under selection, the extent of divergence might be exacerbated because all loci concertedly increase the divergence in that region, potentially forming a large divergence "island" (Feder & Nosil 2010; Flaxman *et al.* 2013). In this case, the effect of recombination in uncoupling neutral and selected loci might be slowed down, but over time still results in narrowed down divergence regions (this will also depend on the strength of selection around the causative loci).

In the samples from the *pseudomajus* x *striatum* hybrid zone, the divergence signal around each candidate gene (*ROS1, ROS2* and *EL-MYB*) is quite sharply defined, with narrow *Fst* "islands" rather than a large island encompassing all loci. This result suggests that recombination in *A. m. pseudomajus* and *A. m. striatum* populations has been ongoing for long enough to break down associations between the selected loci and linked neutral polymorphisms that lie between them. Indeed, between the three main *Fst* peaks in *ROS-EL*, there are several polymorphisms that are shared between samples across the cline (Figure 5.10), suggesting there was sufficient time for recombination to restore some of the diversity around the selected loci.

The effect of recombination in shaping the *Fst* profile around the *ROS-EL* region is not necessarily only due to recombination in the current *pseudomajus* x *striatum* hybrid zone. Recently analysed whole-genome sequence data from seven *A. m. pseudomajus* and nine *A. m. striatum* individuals sampled across the subspecies range (from locations allopatric to the hybrid zone) show a profile of *Fst* remarkably similar to that reported in this work (Annabel Whibley, pers. comm.). The mean level of *Fst* between those samples is higher than between samples near the hybrid zone (likely reflecting their geographic spread), but there are still three prominent peaks around *ROS1, ROS2* and *EL-MYB*. This suggests that the *ROS* and *EL* loci have a long history of divergence between *A. m. pseudomajus* and *A. m. striatum* and that there has been enough time for recombination to finely shape the narrow divergence peaks co-localizing with loci controlling flower colour.

The current pattern of divergence tells us that recombination had a role in shaping the narrow peaks in *ROS* and *EL*. However, it is unclear how the two main haplotypes seen nowadays – *ROS el* and *ros EL* – became established in the first place. I will explore two hypotheses for the origin of these haplotypes and see how either of these might explain the present patterns of divergence. These ideas are not mutually exclusive as both could operate simultaneously or in different times in the history of these populations.

The first hypothesis is that *A. m. pseudomajus* and *A. m. striatum* originate from two distinct lineages, whose geographic ranges did not initially overlap (Figure 7.1A). Let us call these hypothesised ancient lineages the magenta and yellow species. I assume that the two species had functionally similar *ROS-EL* haplotypes as the ones seen nowadays (i.e., magenta is *ROS el* and yellow is *ros EL*). These haplotypes might have become established in each population, either through directional selection or purely by random drift. Assuming that the two lineages split a long time ago, a genomic comparison between the magenta and yellow species should reveal an overall high *Fst* across the whole genome, with no prominent peaks of divergence between them (the whole genome has equally diverged from a common ancestor) (Figure 7.1A-i). If the two species start occurring in parapatry (e.g. due to migrants or range expansion), hybrid zones can form and recombination between the two species' genomes takes place (assuming that the two species are still inter-fertile). In such hybrid zones, loci with no significant effects on fitness are expected to freely exchange between the two species, and thus any differences between them fade away over time (Barton & Gale 1993). This should result in an overall decay of *Fst* across the genome (Figure 7.1A-ii). However, loci under selection (e.g. related to flower colour) will be maintained distinct between the two sides of the hybrid zone, forming sharp clines. As a consequence, these selected loci will constitute divergence peaks in the genome (Figure 7.1A-iii). Due to linkage disequilibrium, these islands should initially be large, but if hybridization is ongoing, they should become increasingly narrowed down (Figure 7.1A-iv). Breaking associations between physically linked loci under selection, such as *ROS* and *EL*, is more difficult as it requires double-recombinant haplotypes to accumulate in the population, leading to a slower decay of *Fst* (the species *ROS-EL* haplotypes are assumed to be maintained together by selection). If this process lasts for many generations, with hybrid zones recurrently forming and collapsing, the final profile of divergence may eventually result in the narrow genomic islands currently seen between *A. m. pseudomajus* and *A. m. striatum* in *ROS-EL* (Figure 7.1C). Making use of the genomic landscape metaphor, this could be dubbed the "eroding plateau" hypothesis, where gene flow between two lineages carrying distinct *ROS el* and *ros EL* plays a fundamental role in establishing the multi-peak profile of *Fst* around *ROS-EL*.

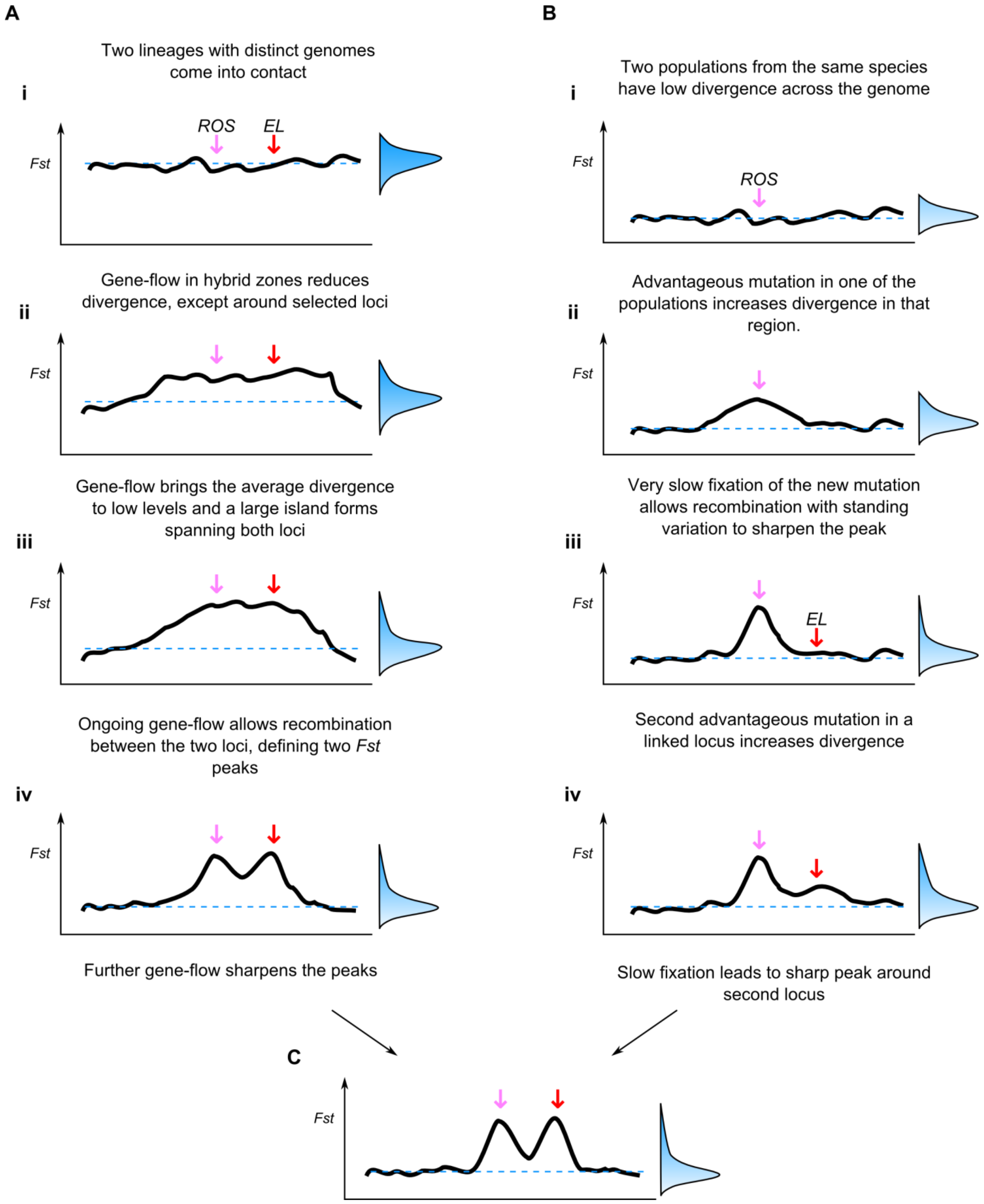

A
i
Two lineages with distinct genomes come into contact
Fst
ROS
EL
ii
Gene-flow in hybrid zones reduces divergence, except around selected loci
Fst
iii
Gene-flow brings the average divergence to low levels and a large island forms spanning both loci
Fst
iv
Ongoing gene-flow allows recombination between the two loci, defining two Fst peaks
Fst
Further gene-flow sharpens the peaks
B
i
Two populations from the same species have low divergence across the genome
Fst
ROS
ii
Advantageous mutation in one of the populations increases divergence in that region.
Fst
iii
Very slow fixation of the new mutation allows recombination with standing variation to sharpen the peak
Fst
EL
iv
Second advantageous mutation in a linked locus increases divergence
Fst
Slow fixation leads to sharp peak around second locus
C
Fst

Figure 7.1 – Hypotheses to explain narrow divergence around *ROS* and *EL* between *A. m. pseudomajus* and *A. m. striatum*.

Each panel represents an imagined *Fst* plot across a genomic region containing two linked loci under selection (equivalent to *ROS* and *EL*). Alongside each *Fst* plot, the distribution of *Fst* across the genome is given. The panels i-iv represent different points in time for each hypothesis.

A) A model considering that the current subspecies of *A. majus* derive from two distinct lineages which came into contact and hybridized. Notice that the mode (dotted line) of the *Fst* distribution becomes progressively lower, but the distribution becomes progressively skewed with a narrow tail of high *Fst* values, corresponding to the loci under selection.

(Figure 7.1 continued)
B) A model considering that the current subspecies of *A. majus* derived from ancestral populations with a shared allelic pool. Two successive selective sweeps in *ROS* and *EL* create the two individual peaks. The peaks become sharp because fixation of the new allele is slow, allowing enough time for recombination within the population to sharpen the divergence around the locus. Notice that the mode (dotted line) of the *Fst* distribution remains low, but the distribution gets progressively skewed due to the fixation of loci that undergo successive selective sweeps.
C) A schematic of the current multi-peak *Fst* observed around *ROS* and *EL*.

An alternative hypothesis is that the modern *A. m. pseudomajus* and *A. m. striatum* originate from a single lineage, within which successive selective sweeps led to the fixation of new functional alleles in *ROS* and *EL* (Figure 7.1B). Assume that there was an ancient magenta (<u>*ROS el*</u>) species, but a mutation appears in a particular population that creates a new <u>*ros el*</u> haplotype. In this case, the *Fst* between the population with the new *ros* allele and a population carrying the *ROS* allele will be overall low (they are effectively still the same species) (Figure 7.1B-i). If the new haplotype is locally advantageous, it will start to increase in frequency in that population, increasing divergence in the surrounding region of this locus (Figure 7.1B-ii). However, being a recessive haplotype, its fixation in the population is slow (the heterozygotes are not at an advantage), and so there is opportunity for recombination to occur and narrow down the divergence around the new *ros* allele (Figure 7.1B-iii). In this new <u>*ros el*</u> background a mutation that changes *el* → *EL* is now advantageous, and it too becomes fixed by a similar process. This creates a second *Fst* peak, which appears neighbouring the firstly established peak in the *ROS* locus (Figure 7.1B-iv). Extending the landscape metaphor, this situation could be dubbed as the “rising peaks” hypothesis.

The two explanations ("eroding plateau" or "rising peaks") are not different in that both illustrate the importance of gene flow and recombination in sharpening the peaks of *Fst*. Whether the modern *ROS-EL* haplotypes in *A. m. pseudomajus* and *A. m. striatum* originated from initially allopatric species ("eroding plateau" hypothesis) or in parapatric populations of the same species ("rising peaks"), it is undeniable that the two subspecies hybridize nowadays. Therefore, current and past hybridization events have likely played a significant role in shaping the divergence across the genome between *A. m. pseudomajus* and *A. m. striatum*. In particular, the present study shows that hybrid zones between these subspecies can persist long enough to build-up recombinants between tightly linked loci under selection, such as *ROS* and *EL*. Over time, this recombination might establish shared polymorphisms between the subspecies, narrowing down the *Fst* peaks to tightly coincide with the selected loci.

The two hypotheses presented here can be more rigorously addressed by performing computer simulations of divergent genomes recombining in a hybrid zone (Nick Barton, pers. comm.). Such simulations are underway, and should allow exploration of the different initial scenarios (allopatry vs. parapatry) and conditions (e.g. rate of gene flow, number of generations, strength of selection) that may be required to obtain a profile of divergence similar to the one observed for *ROS-EL*.

Ultimately, understanding the current divergence patterns between *A. m. pseudomajus* and *A. m. striatum* will require knowing more about the origin of their haplotypes. This is particularly relevant in this system, which comprises around 20 species of *Antirrhinum* in Europe, some of them with *ROS-EL* haplotypes which are functionally identical to the *A. m. pseudomajus* or *A. m. striatum* haplotypes (Schwinn *et al.* 2006; Whibley *et al.* 2006). Moreover, these different species are inter-fertile with each other in glasshouse crosses, raising the possibility that the *ROS-EL* haplotypes in *A. m. pseudomajus* and *A. m. striatum* originated through past introgression from other *Antirrhinum* species (supporting the "eroding plateau" hypothesis). This is not an unlikely scenario, as exemplified by the well-studied pervasive introgression of loci controlling wing colour patterns in *Heliconius* butterflies, through a long history of current and past hybridization events (Pardo-Diaz *et al.* 2012; The Heliconius Genome Consortium 2012; Brower 2013; Martin *et al.* 2013). The access to genome-wide data from multiple species of *Antirrhinum* should help address some of these questions in the future.

7.2 Selection on flower colour may be related to the pollination syndrome of *Antirrhinum*

Despite the unknown evolutionary origin of the two functional *ROS-EL* haplotypes in *A. m. pseudomajus* and *A. m. striatum*, it seems plausible that they did not evolve stochastically (i.e. without a role for selection) to produce the floral phenotype that characterizes the two subspecies. Instead, it is reasonable to speculate that their evolution is related to the overall adaptation of *Antirrhinum* flowers to attract their restricted group of insect pollinators, mostly composed of large bee species (Tastard *et al.* 2008; Vargas *et al.* 2010).

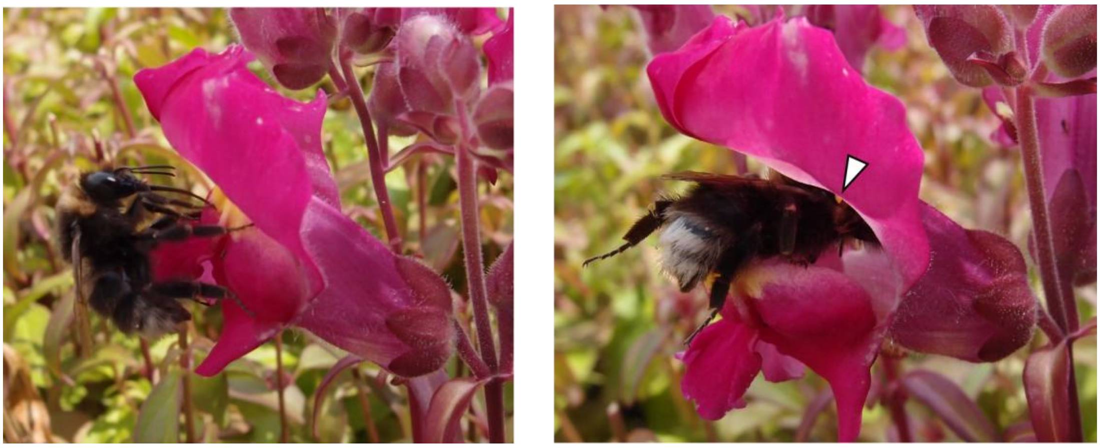


Figure 7.2 – A bumblebee pollinating an *Antirhrinum majus* flower.
The two photographs are successive snapshots of a bumblebee entering the flower to access the nectar, stored in the lower part of the flower's tube. Several key features of the pollination syndrome in *Antirrhinum* are evident. The pollen and nectar are stored within the closed flower, allowing its access to large bee species that are strong enough to open it. The bent shape of the ventral petal of *Antirrhinum* serves as a "landing platform" for the bee (image on the left). As the bee enters the flower (on the right), its back rubs against the pollen-containing anthers and the pistil, both located on the dorsal part of the tube (a visible anther is indicated by the white arrow head). Notice that self-fertilization in wild *Antirrhinum* is avoided due to a physiological self-incompatibility system (Xue *et al.* 1996).

Antirrhinum flowers constitute a clear example of a pollination syndrome, that is, they encompass a set of traits associated with a specific group of animal pollinators (Fenster *et al.* 2004). This set of traits includes: the flowers' bilateral symmetry, with a closed corolla which hinders access to rewards such as pollen and nectar by smaller animals; the presence of a landing platform for the bees, provided by their folded ventral petal; the accumulation of nectar in the lower part of the corolla tube, which requires the insects to enter deep in the flower to harvest that reward; the position of the anthers in the dorsal part of the flower tube, which allows depositing pollen in the back of the insect, where it is more difficult for it to remove it by grooming; and the positioning of the pistil near the anthers, which allows deposition of compatible pollen that the bee might carry from another plant

(Whitney & Glover 2007) (Figure 7.2). With regards to flower colour in particular, several experiments involving inbred mutant lines of *A. majus* have shown that this trait significantly influences the visitation behaviour of bumblebees to different flowers (Glover & Martin 1998; Niovi Jones & Reithel 2001; Shang *et al.* 2011; Whitney *et al.* 2013).

Flower colour is primarily used by bee pollinators as a means to distinguish a flower from the green and brownish background of vegetation's foliage and soil, which looks more or less homogeneous in the insect's visual spectrum (Chittka & Raine 2006). Bees have trichromatic vision, being able to perceive light reflectance in the green, blue and UV spectrums. This means that both a magenta flower (such as *A. m. pseudomajus*) and a yellow flower (such as *A. m. striatum*) contrast with the background. Indeed, field trials using a magenta line of *A. majus* (similar to *JI7*) and a yellow line (a *sulf* mutant, similar to *A. m. striatum* flowers) showed that both types of flowers are equally visited by pollinators in a mixed array of phenotypes (Niovi Jones & Reithel 2001). On the other hand, white or ivory *A. majus* flowers (*niv* and *rosea*^{dorsea} mutants, respectively) are not visited as much as magenta flowers in mixed arrays, possibly because they are less well distinguished from the background (Glover & Martin 1998; Dyer *et al.* 2007). However, if a dominant allele of *VEN* (conferring vein-specific pigmentation), is introduced in the *rosea*^{dorsea} background, then flowers are visited as frequently as full magenta flowers (Shang *et al.* 2011; Whitney *et al.* 2013). The venation pattern is thought to serve as a nectar guide in *Antirrhinum*, that is, a visual cue that diminishes the handling time necessary for the pollinator to enter the flower and access its reward.

The two subspecies used in this work, *A. m. pseudomajus* and *A. m. striatum*, have strikingly distinct flower colours. The two main pigments produced in these flowers – the magenta-coloured anthocyanins and the yellow-coloured aurones – form two visually contrasting colour patterns in the subspecies. *A. m. pseudomajus* has magenta pigmentation that spreads throughout most of the corolla, except in the ridges where the ventral and lateral petals fuse (known as the "foci"), where it is mostly absent (Figure 7.3). In its place there is a visible yellow patch that co-localizes with the insects' main entrance point to access the inside of the flowers (where the nectar and pollen are found). As an almost mirror image of this, *A. m. striatum* has spread yellow pigmentation in the corolla, with magenta pigmentation only in the petal's veins, which is restricted to the junction of the dorsal petals, again near the insects' access point to the inside of the flower (Figure 7.3).

These coordinated patterns, formed by two colours that bees can distinguish well, are likely to efficiently contrast the flowers from the landscape background at a distance. Furthermore, they might also create a specific contrast that highlights the pollinators' entrance point to the flower, providing a nectar guide at a closer distance. Testing these ideas will require specifically analysing *A. m. striatum* and *A. m. pseudomajus* plants in pollination tests under controlled conditions. Although several pollination experiments with *Antirrhinum* have been undertaken, none of them have used the wild subspecies focused on in this work. The genetic dissection of flower colour from this and other works, should enable the mechanism by which pollinators learn to discriminate and handle flowers with well-characterised genotypes/phenotypes to be investigated. Particularly related to the present work, it will be important to find out if the phenotypes conferred by different *ROS-EL* haplotypes (namely recombinant ones) are perceived and/or handled differently by pollinators; for example, can they distinguish restricted (*ros EL*) from spread (*ros el*) venation?

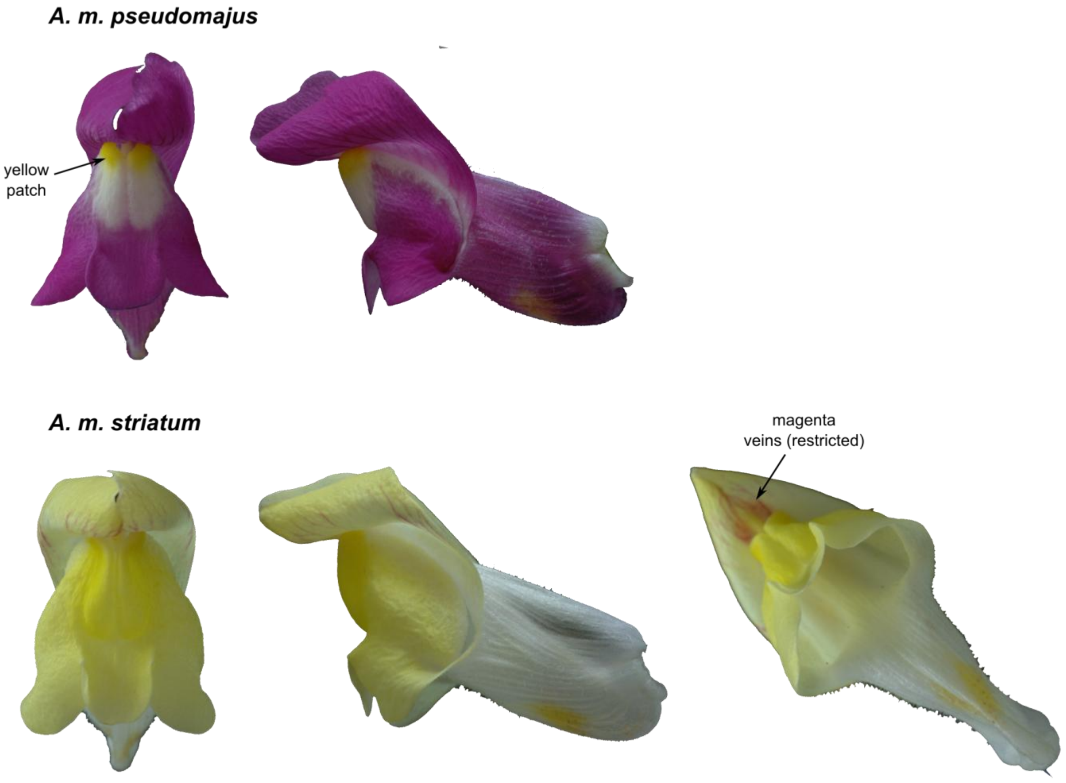


Figure 7.3 – Typical flowers of *A. m. pseudomajus* and *A. m. striatum*.
Photographs show the typical colour pattern of the two subspecies. *A. m. pseudomajus* is characterized by spread magenta pigmentation and a yellow patch restricted to the ridges ("foci") where the ventral and lateral petals meet. *A. m. striatum* is characterized by spread yellow pigmentation and magenta venation, which is intensified in the region where the two dorsal petals meet. Individuals in the photographs are derived from populations located further than 4 km from the *pseudomajus* x *striatum* hybrid zone centre (magenta individual is V164-60; yellow individual is V206-40).

Although I focus on the role of pollinators in the evolution of the flower colour patterns of *A. m. pseudomajus* and *A. m. striatum*, anthocyanin pigments in particular are often involved in the adaptation to other biotic and abitotic conditions, such as herbivory resistance, drought stress, heat stress and soil moisture (reviewed in Chalker-Scott 1999; Strauss and Whitall 2006). In the *pseudomajus* x *striatum* hybrid zone focused on in this work, relevant ecological factors that differ across the geographic transect may be unaccounted for (Khimoun *et al.* 2012). Therefore, other agents of selection, besides pollinators, may explain the flower colour divergence in these subspecies. It would be desirable to perform a systematic characterization of some of the main features of the environment, at a sufficient geographic resolution, to characterize the current hybrid zone transect (e.g. mean temperature, soil moisture, soil pH, sun exposure, flora composition, etc.) and see if they are likely to explain the maintenance of the observed sharp cline for flower colour.

Although ecological factors that are unaccounted for may contribute to the sharp cline between the two subspecies, there seems to be a selective pressure specifically related to the flower colour patterns in *Antirrhinum*. Three arguments support this idea. First, *ROS-EL* controls anthocyanin production mainly in the flowers, whereas the pigmentation in other organs, such as leaves, must be controlled by other loci (for example, leaf anthocyanin pigmentation is visible in *rosea*^{dorsea} and *A. m. striatum* plants). Second, the phenotypes of *A. m. pseudomajus* and *A. m. striatum* are not merely related to the presence or absence of pigments, but confer a particular pattern which is interwoven with the morphological architecture of the flowers (notice, for example, the absence of magenta pigment where the yellow patch occurs in *A. m. pseudomajus* flowers; Figure 7.3). Finally, the floral phenotype is not just characterized by the magenta anthocyanins, but also by the yellow pattern of aurones, and these two components are likely to interact at a fitness level (Whibley *et al.* 2006). For example, out of the around 20 species of *Antirrhinum* in Europe, none of them have completely white flowers (they always have venation) nor orange flowers (resulting from the overlap of anthocyanins and aurones), despite it being possible to find such phenotypes in the *pseudomajus* x *striatum* hybrid zone. This suggests that the pattern of flower colour in these subspecies is an integrated feature likely to be under selection. The most parsimonious explanation for this observation is the relation of flower colour with the pollination ecology of this species.

In conclusion, it seems reasonable to speculate that the patterns of pigmentation in *A. m. pseudomajus* and *A. m. striatum* occur as two fit phenotypes that are maintained by selection in the populations flanking the hybrid zone between these subspecies. Seen from this perspective, the loci that control different aspects of the floral phenotype may be seen as co-adapted complexes that interact to produce the fit phenotypes. In particular, the *ROS-EL* region, which contains several loci controlling the magenta pigment of the flowers, might produce more or less fit combinations depending on the allelic combination across those loci.

7.3 Modes of selection maintaining *ROS-EL* haplotypes

I have shown that *ROS* and *EL* show sharp allelic clines in the *pseudomajus* x *striatum* hybrid zone, suggesting that the subspecies *ROS el* and *ros EL* haplotypes are impeded to be freely exchanged between the two populations. This suggests that hybrid phenotypes incur a fitness cost, or else the clines would have faded away over time (Barton & Gale 1993). Importantly, the magenta phenotype conferred by the genotype in one locus, depends on the genotype in the other locus (e.g. Figure 6.2), that is, the two loci interact at a genetic level in regulating this trait. Recombinant haplotypes of *ROS-EL* occur in the *pseudomajus* x *striatum* hybrid zone at relatively high frequencies in the central region of the geographic transect (~5%). However, they are rare in the flanks of the hybrid zone, suggesting that recombinants might be impeded from introgressing to the parental populations due to selection (although this could also be explained by an insufficient time for them to spread). Moreover, these recombinant haplotypes were shown to significantly alter the phenotype of the flowers, which could have consequences for the fitness of individuals. Two modes of selection were proposed to account for the maintenance of the *ROS el* and *ros EL* haplotypes in the two *A. majus* subspecies: frequency-dependent selection and co-adaptation between loci.

7.3.1 Frequency-dependent selection

Selection on flower colour may act in a frequency-dependent manner, particularly if we consider the constancy of pollinator's behaviour to visit similarly coloured flowers during each foraging bout (discussed in section 6.3.3). Under this view, the more common phenotype may be preferred over rarer phenotypes (positive frequency-dependent selection), which can make selection particularly strong in the flanks of the hybrid zone. For example, an individual with magenta phenotype may have a fitness of 1 in the *A. m. pseudomajus* flank of the hybrid zone, but its fitness drops (<< 1) if it occurs in the *A. m. striatum* flank. In the case of frequency-dependent selection, the fitness of an individual changes depending on its surrounding context. This means that in the *pseudomajus* x *striatum* hybrid zone, a recombinant haplotype is not intrinsically less fit than non-recombinant haplotypes; it just happens to occur at a low frequency and therefore cannot become established in the population.

Interestingly, if a trait is under positive frequency-dependent selection, polymorphisms in the population will not be maintained in the long term (Ridley 2004). This is because, as soon as one phenotype becomes the commonest, it will be the fittest and therefore the genetic combination that produces that phenotype will become fixed by selection in the population. Which haplotype becomes fixed might simply be determined by chance and, therefore, two populations might become fixed for different phenotypes. If those two populations then come into contact in a hybrid zone, a reproductive incompatibility arises because the two phenotypes are locally more frequent in the flanks of the hybrid zone. This creates a barrier to the introgression of alleles controlling the trait from one population to the other, and a clinal hybrid zone might form. Clines can be stably maintained by positive frequency-dependent selection because genotypes that disperse along the cline will be selected against if they produce one of the rarer phenotypes in the area (Mallet & Barton 1989). Regardless of how the *A. m. pseudomajus* and *A. m. striatum* flower colour phenotypes became established in the first place (i.e., whether it involved pollinator-mediated directional selection or not), the fact that they are nowadays the two most common phenotypes in the populations across the Pyrenees, will make it difficult for any other phenotype to become established under positive frequency-dependent selection. In this sense, the two parental *ROS el* and *ros EL* haplotypes will be, on average, fitter than any recombinant haplotype, because the latter will always be the rarest in the population. Therefore, positive frequency-dependent selection may explain the maintenance of fit allelic combinations across multiple loci in a hybrid zone.

7.3.2 Co-adaptation

ROS and *EL* might also be hypothesised to constitute a co-adapted gene complex, where fit allelic combinations from *A. m. pseudomajus* and *A. m. striatum* are maintained by selection, despite current and past gene flow events between these subspecies (Whibley *et al.* 2006; Khimoun *et al.* 2011). The term co-adaptation refers to cases where the genotype in one locus is differently favoured by selection depending on the genotype in another locus (or loci) (Dobzhansky 1950; Wallace 1991). In other words, the fitness advantage of an allele in a locus does not simply add up to the total fitness of an individual, but rather depends on the alleles that that individual carries in other loci in the genome (this is a definition of fitness epistasis). The co-adapted nature of multiple loci in a genome plays a fundamental role in the evolution of reproductive barriers between populations, as incompatibilities are established not by a single gene, but by multiple genes that interact with each other in a given genomic context (Phillips 2008). In this context, does *ROS-EL* constitute a co-adapted gene complex?

7.3.2.1 A genotype-phenotype graph for ROS-EL haplotypes

To understand how the genetic epistasis between *ROS* and *EL* may translate to fitness epistasis (and thus, offer the conditions that support a co-adaptation hypothesis) I will explore different scenarios for the interplay between genotype, phenotype and fitness. To aid in presenting these ideas I will represent *ROS-EL* haplotypes in a simplified fitness landscape (also known as a fitness graph; Crona et al. 2013) (Figure 7.4). The concept of fitness landscape was originally presented as a metaphor to explain how the fitness can change in a genotypic space where epistatis takes place (Wright 1932). The metaphor draws a comparison with a geographic landscape (composed of mountains and valleys): the horizontal plane of a fitness landscape is used to represent the genotypes of individuals (or sometimes allele frequencies in a population) and the height of the landscape corresponds to the fitness of those genotypes. Therefore, peaks in the mountainous fitness landscape represent adaptive genotypes, whereas valleys represent maladaptive ones. This concept has been vastly expanded to incorporate formal theoretical formulations (reviewed in Gavrilets 2010), but I will use it in its simplified form, as a framework to qualitatively explore different selective scenarios for both epistatic and non-epistatic fitness interactions between *ROS* and *EL*.

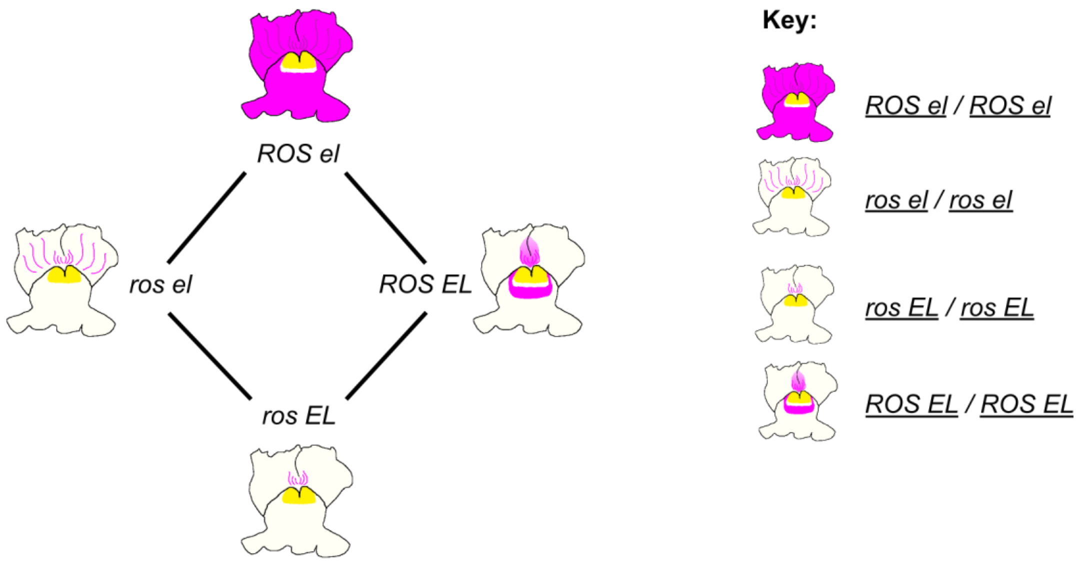


Figure 7.4 – A genotype-phenotype graph between *ROS-EL* haplotypes and flower colour phenotype.
Each cartoon represents the expected phenotype conferred by each of the four *ROS-EL* haplotypes in a homozygous state. Neighbouring genotype-phenotype combinations are separated by a change in a single locus.

Before making considerations about fitness, a simple genotype-phenotype graph can be built based on the gained knowledge of how the genotype in *ROS* and *EL* affects the magenta phenotype of the flowers (Figure 7.4). Let us consider four phenotypes, corresponding to each of the four *ROS-EL* haplotypes in a homozygous state and with *VEN* in the background: magenta with cryptic spread veins (<u>*ROS el*</u>/<u>*ROS el*</u> = *A. m. pseudomajus*); Eluta with restricted veins (<u>*ROS EL*</u>/<u>*ROS EL*</u>); non-magenta with visibly spread veins (<u>*ros el*</u>/<u>*ros el*</u>); and non-magenta with visibly restricted veins (<u>*ros EL*</u> /<u>*ros EL*</u> = *A. m. striatum*). In this simple two-locus, two-allele genotype-phenotype graph, each haplotype is separated from another by an allelic change in one locus at a time (Figure 7.4). Under the assumption that fitness is correlated to the floral phenotype, we can explore different scenarios for selection both with and without fitness epistasis.

7.3.2.2 Exploring fitness graphs for ROS-EL haplotypes

One hypothesis to explain the maintenance of the *ROS el* and *ros EL* haplotypes in *A. majus* subspecies is to assume that the phenotype of each subspecies has equal fitness, which is greater than that of the recombinant phenotypes (Figure 7.5A). Under this assumption, any step changes in the landscape away from the parental *ROS el* or *ros el* haplotypes involve a fitness cost to the individuals. This landscape is epistatic, since allelic changes in each locus have different fitness consequences depending on the genotype of the other locus. For example, a change from *ROS* → *ros* has a fitness cost when changing from *ROS el* → *ros el*, but a fitness gain when the change is from *ROS EL* → *ros EL* (Figure 7.5A). The magnitude of the fitness change may or may not be the same in each case (in the examples given in Figure 7.5 I consider different magnitudes). The crucial point is that the same allelic change in one locus results in a fitness change with a different sign (a gain or a loss), depending on the allele at the other locus.

We can also conceive that the phenotype of one of the *ROS-EL* recombinants is as fit as either subspecies phenotype. For example, let us consider that the pattern of venation is not under selection (Figure 7.5B). In that case homozygotes *ros el* / *ros el* have similar fitness to those with the *striatum* *ros EL* / *ros EL* genotype. This landscape is also epistatic, because the same allelic change does not always result in the same fitness change. For example, a change from *ros* → *ROS* is neutral (no fitness gain or loss) when changing from *ros el* → *ROS el*, but confers a fitness cost when changing from *ros EL* → *ROS EL* (Figure 7.5B).

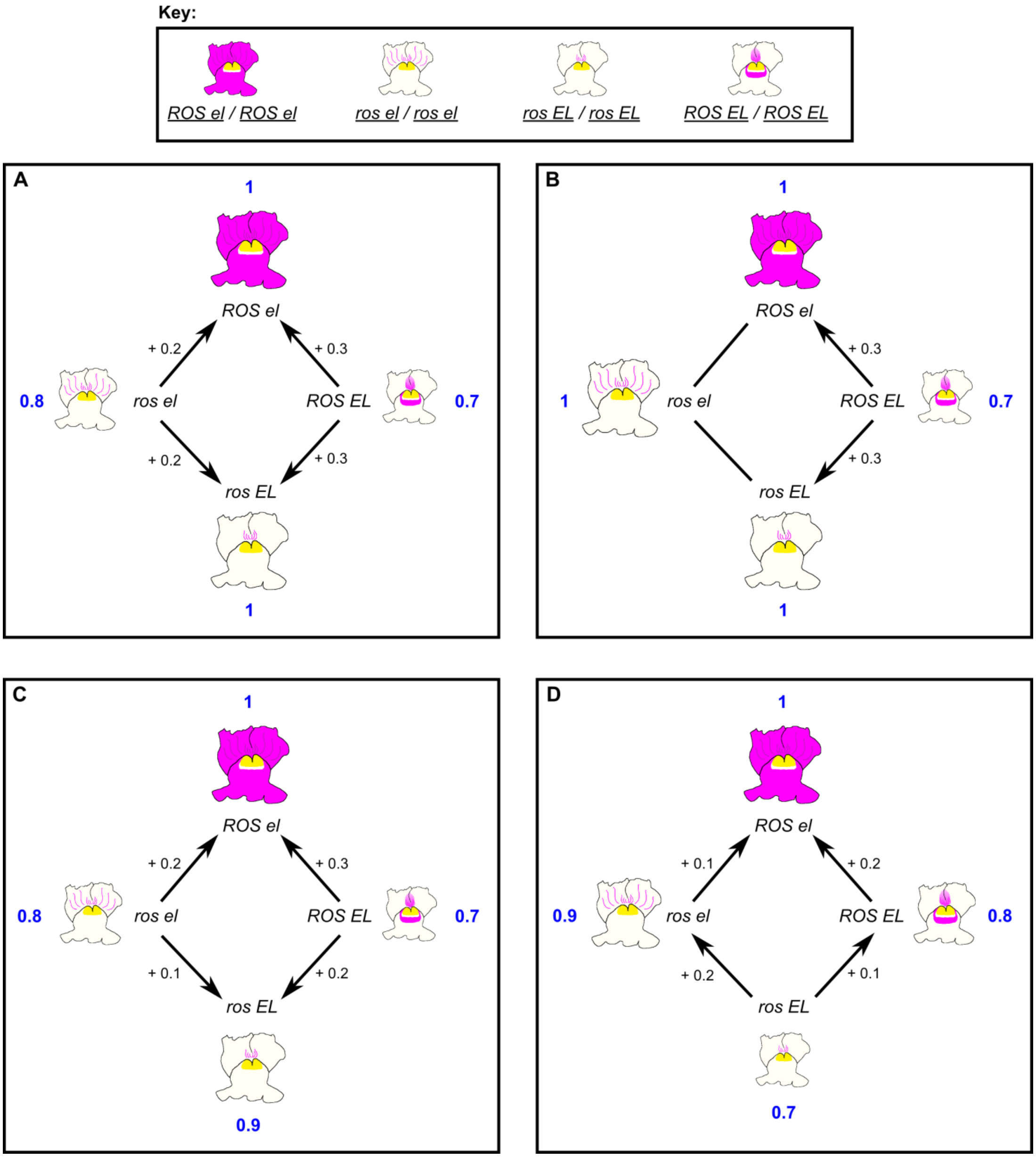


Figure 7.5 – Example fitness graphs for *ROS-EL* haplotypes.

Legend as in Figure 7.4, except that fitness values (blue numbers) are indicated for each of the phenotypes. The size of each pictogram is proportional to the fitness. Arrows point to allelic changes that result in a fitness gain (changes in the opposite direction are assumed to incur a fitness cost of the same magnitude). The fitness values are not based on empirical data, instead they are given as examples to accompany the interpretation of the graphs.

A) An epistatic fitness landscape where each of the subspecies phenotypes is equally fit (fitness = 1), and fitter than both recombinant phenotypes (fitness < 1). The landscape is epistatic, since the same allelic change in one locus confers a fitness gain or a fitness cost, depending on the genotype at the other locus. For example, changes from *ros* → *ROS* and from *EL* → *el* have a fitness cost when changing away from the *ros EL* haplotype, but a fitness gain when moving towards a *ROS el* haplotype.

(Figure 7.5 continued)

B) An epistatic fitness landscape where one of the recombinant phenotypes is less fit (fitness < 1) than all other phenotypes, which are all equally fit (fitness = 1). The landscape is epistatic, since the same allelic change in one locus is either neutral or not, depending on the genotype of the other locus.

C) An epistatic fitness landscape where one of the phenotypes is fitter than all others. This case is similar to Figure 7.5A, except that one of the subspecies phenotypes (magenta, *ROS el* / *ROS el*) is considered the fittest. However, allelic changes in the two loci do not always confer the same fitness change, therefore the landscape is epistatic.

D) An additive fitness landscape. Under the assumption of additive fitness, one of the phenotypes has to be fitter than the others. In this case, the magenta phenotype is assumed to be the fittest, and any steps away from it incur a fitness cost. In this landscape, allelic changes in each locus always confer the same fitness gain or cost, independently from the genotype in the other locus. In the example, a change from *ros* → *ROS* always confers a fitness gain of +0.1 and a change from *EL* → *el* a fitness gain of +0.2.

An epistatic fitness landscape can also accommodate cases where one phenotype is fitter than all others (Figure 7.5C). For example, if the phenotype of one of the subspecies is fitter than the other subspecies (in this example *A. m. pseudomajus* has higher fitness than *A. m. striatum*), but both subspecies are still fitter than either recombinant, and therefore the allelic changes in each locus are not additive with regards to fitness.

Finally, an alternative to fitness epistasis is when fitness is additive and therefore the loci cannot be considered to be co-adapted. Under an additive fitness model of selection, it is assumed that changes from one allele to another in a particular locus result in a constant change in fitness, regardless of the genotype in other loci (Phillips 2008; Crona *et al.* 2013). As a consequence of this additive property, one particular phenotype is, by definition, fitter than all others (the phenotype which results from combining all the alleles that confer a fitness gain). Applied to the case of flower colour in *Antirrhinum*, if the fitness conferred by particular alleles in *ROS* and *EL* is additive, then *A. m. pseudomajus* and *A. m. striatum* phenotypes cannot have an equal fitness (Figure 7.5D). For example, consider there is a fitness gain by changing *ros*→ *ROS* and by changing *EL* → *el* (the magnitude of the fitness change does not have to be the same for each locus) (Figure 7.5D). In this scenario, the *ros EL* haplotype is the less fit, *ROS EL* and *ros el* haplotypes have intermediate fitness (not necessarily the same) and the *ROS el* haplotype has the highest fitness, which results from the summation of the fitness gains in each allelic step-change (Figure 7.5D). Under this scenario, even though *ROS* and *EL* interact epistatically at the genetic level, this is not reflected as fitness epistasis.

By qualitatively exploring different selective scenarios for *ROS-EL* haplotypes, it becomes clear that the loci can be considered co-adapted (i.e. there is fitness epistasis) as long as the two subspecies phenotypes confer higher fitness than the phenotype of at least one of the recombinants (Figure 7.5A-C). Strictly speaking, this work does not directly test the hypothesis that *ROS-EL* occur as a co-adapted gene complex. This would involve measuring the fitness of individual phenotypes, by determining if there are differences in the reproductive success of the individuals in the population. However, some empirical evidence supports against its alternative that the fitness conferred by *ROS* and *EL* genotypes is simply additive.

If the fitness conferred by *ROS* and *EL* was additive, one haplotype should be fitter than all others. In that case, a stable hybrid zone is not expected to be maintained for a long time, since the universally fit haplotype would easily take over the population and the hybrid zone would collapse (Barton & Gale 1993). Contrary to this, in the *pseudomajus* x *striatum* hybrid zone detailed in this work, both *ROS1* and *EL-MYB* have sharp allelic clines, and occur in high linkage disequilibrium in the core of the hybrid zone. Therefore, it seems unlikely that one of the subspecies haplotypes (*ROS el* or *ros EL*) is much fitter than the other, otherwise shallow clines with long tails of introgression would be expected to occur.

I have, however, found a weak signal of introgression of *A. m. pseudomajus ROS1*^{p} and *el-myb*^{p} alleles towards the *A. m. striatum* side of the hybrid zone (Figure 6.5B), which could support the hypothesis that the *A. m. pseudomajus* haplotype is fitter than *A. m. striatum* haplotype. However, this result has to be interpreted with care, since it might not be related to selection, but rather to population structure. Because hybrid zones are maintained by a balance between selection and dispersal, they can move from one place to another over time (Barton & Hewitt 1985; Barton & Gale 1993). This movement might depend on the local density of individuals, which is likely to be heterogeneous in nature (and may change over the years by chance). For example, if there is a difference on the population sizes on either side of the hybrid zone, there will be an unbalanced input of alleles to the hybrid zone centre. The population with higher density and dispersal may thus spread, despite there being no fitness differences between individuals from the two populations. In some cases, these fluctuations may result in the extinction of a hybrid zone, with one of the populations eventually taking over the other. Therefore, the introgression of alleles into the *A. m. striatum* flank of the hybrid zone may be related to other factors other than selection (this can be investigated in more detail in the future, for example, by analysing the density of individuals across the geographic transect of the hybrid zone).

There is strong evidence that several hybrid zones formed between these two subspecies in the past (Khimoun *et al.* 2011). These extinct hybrid zones were detected due to a discrepancy between the genotype of a subspecies-specific chloroplast marker and the flower colour phenotype of a population. Allopatric populations of *A. m. pseudomajus* are fixed for a particular allele of this chloroplast marker, whereas allopatric populations of *A. m. striatum* are fixed for a different allele. However, at the edges of each subspecies distribution range, some *A. m. pseudomajus* populations were fixed for an *A. m. striatum* chloroplast marker, and in other populations the opposite was observed. This suggests that hybrid zones formed in the past, and the flower colour alleles of one of the subspecies invaded the other. Crucially, though, the subspecies' alleles that become fixed are not always the same, further support against one of the haplotypes being generally fitter than the other. Incidentally, these data are not incompatible with one of the haplotypes being fitter than another under certain environmental conditions (Khimoun *et al.* 2012), but this is still compatible with the hypothesis that they are co-adapted.

In summary, it seems plausible that the subspecies allelic combinations of *ROS-EL* have been maintained by selection (through co-adaptation and/or positive frequency-dependent selection), despite repeated opportunities for recombination through hybridization. Formally proving co-adaptation is not trivial, as it requires measuring individuals’ fitness in the population, which is technically challenging. One approach that is undergoing in the current *pseudomajus* x *striatum* hybrid zone is to construct a multi-generation pedigree at the population level, which should allow us to determine how many progeny each individual contributes to the next generation (a direct measure of its fitness). This should provide evidence for whether certain genotypes/phenotypes for flower colour are intrinsically fitter than others (co-adaptation), and test hypotheses about frequency-dependent forms of selection.

8 Concluding Remarks

This work shows that isolating barriers between populations can involve a few loci of major effect, fitting with a "genic" rather than "genomic" view of divergence (Wu 2001; Lexer & Widmer 2008). When gene flow occurs between populations (e.g. in hybrid zones), those barrier loci will form regions of higher divergence compared with the rest of the genome. Recent years have seen a burst in reported cases where these heterogeneous patterns of genomic divergence are observed between hybridizing populations, much of them driven by the application of high-throughput sequencing methods to survey natural populations. This provides a "panoramic view" of the genomic divergence between hybridizing populations; for example: how many regions of high divergence are there across the genome? How large are those regions? Are they gene-rich? Are they associated with particular chromosomal regions (e.g. inversions, centromeres, telomeres, etc...)? Despite being useful, this "panorama" says little about the actual genetic basis of isolating barriers and so, the divergence signals are difficult to interpret because it is unknown how many (if any) functional loci lie within them.

This work aimed at partially filling in this gap, by combining genomic, genetic and population genetics approaches to characterize a region containing several loci controlling a character under selection in a hybrid zone. The genetic dissection of loci within the *ROS-EL* region proved invaluable in interpreting the divergence signals in this region, which are quite heterogeneous, despite the tight linkage between the selected loci. This illustrates the consequences of past and present selection, gene flow and recombination in shaping divergence patterns at a fine-scale. Finally, this work emphasises how genetic interactions (epistasis) might lead to the maintenance of fit genotypes despite gene flow, because unfit phenotypes may be produced by allelic combinations that uncouple the parental alleles.

The work presented here for *ROS-EL* can be extended to other loci in the genome. Indeed, flower colour is unlikely to be the only trait under selection in the hybrid zone: other regions of high divergence occur across the genome, which currently have unknown function, but are not clearly associated with flower colour (Louis Boell, pers. comm.). Studying these loci both at the population and genetic levels will no doubt provide an integrated view of how divergence in the genome relates with the genetic basis of isolating barriers between populations.

9 References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of evolutionary biology*, **26**, 229–46.

Aharoni A, De Vos CH, Wein M *et al.* (2001) The strawberry FaMYB1 transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tobacco. *Plant Journal*, **28**, 319–32.

Alberts B, Johnson A, Lewis J *et al.* (2002) How Cells Read the Genome: From DNA to Protein. In: *Molecular Biology of the Cell* 4th Edition, pp. 302–335. Garland Science, London.

Anderson EC, Skaug HJ, Barshis DJ (2014) Next-generation sequencing for molecular ecology: A caveat regarding pooled samples. *Molecular ecology*, **23**, 502–12.

Angert AL, Schemske DW (2005) The evolution of species' distributions: reciprocal transplants across the elevation ranges of *Mimulus cardinalis* and *M. lewisii*. *Evolution*, **59**, 1671–84.

Barton NH, Gale KS (1993) Genetic Analysis of Hybrid Zones. In: *Hybrid Zones and the Evolutionary Process* (ed Harrison RG), pp. 13–45. Oxford University Press.

Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–48.

Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.

Bateson W, Saunders E, Punnett R (1905) Report II. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society*.

Baur E (1911) Ein Fall von Faktorenkoppelung bei *Antirrhinum majus*. *Verhandlungen des naturforschenden Vereines in Brünn*, **XLIX**, 130–38.

Baur E (1912) Vererbungs- und Bastardierungsversuche mit *Antirrhinum* - II. Faktorenkoppelung. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, **VI**, 34–98.

Baxter SW, Nadeau NJ, Maroja LS *et al.* (2010) Genomic Hotspots for adaptation: the population genetics of Mullerian mimicry in the *Heliconius melpomene* clade. *Plos Genetics*, **6**, e1000794.

Baxter SW, Papa R, Chamberlain N *et al.* (2008) Convergent evolution in the genetic basis of Müllerian mimicry in *Heliconius* butterflies. *Genetics*, **180**, 1567–77.

Beaumont MA (2005) Adaptation and speciation: what can Fst tell us? *Trends in Ecology & Evolution*, **20**, 435–40.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: Models and empirical evidence. *Annual Review of Ecology Evolution and Systematics*, **38**, 459–87.

Bomblies K, Lempe J, Epple P *et al.* (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biology*, **5**, e236.

Brower AVZ (2013) Introgression of wing pattern alleles and speciation via homoploid hybridization in *Heliconius* butterflies: a review of evidence from the genome. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20122302.

Buerkle CA, Lexer C (2008) Admixture as the basis for genetic mapping. *Trends in ecology & evolution*, **23**, 686–94.

Burr FA, Burr B, Scheffler BE *et al.* (1996) The maize repressor-like gene intensifier1 shares homology with the r1/b1 multigene family of transcription factors and exhibits missplicing. *The Plant cell*, **8**, 1249–59.

Carroll SB (2005) Evolution at two levels: on genes and form. *PLoS biology*, **3**, e245.

Causier B, Castillo R, Xue Y, Schwarz-Sommer Z, Davies B (2010) Tracing the evolution of the floral homeotic B- and C-function genes through genome synteny. *Molecular biology and evolution*, **27**, 2651–64.

Chalker-Scott L (1999) Environmental significance of anthocyanins in plant stress responses. *Photochemistry and Photobiology*, **70**, 1–9.

Chittka L, Raine NE (2006) Recognition of flowers by pollinators. *Current Opinion in Plant Biology*, **9**, 428–35.

Chittka L, Thomson JD (1997) Sensori-motor learning and its relevance for task specialization in bumble bees. *Behavioral Ecology and Sociobiology*, **41**, 385–98.

Clarkson CS, Weetman D, Essandoh J *et al.* (2014) Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nature communications*, **5**, 4248.

Cooley AM, Modliszewski JL, Rommel ML, Willis JH (2011) Gene duplication in *Mimulus* underlies parallel floral evolution via independent trans-regulatory changes. *Current biology*, **21**, 700–4.

Counterman BA, Araujo-Perez F, Hines HM *et al.* (2010) Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in *Heliconius erato*. *PLoS Genetics*, **6**, e1000796.

Crawford JE, Nielsen R (2013) Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping? *Molecular ecology*, **22**, 6131–48.

Crona K, Greene D, Barlow M (2013) The peaks and geometry of fitness landscapes. *Journal of theoretical biology*, **317**, 1–10.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology*, **23**, 3133–57.

Cruzan MB, Arnold ML (1994) Assortative mating and natural selection in an *Iris* hybrid zone. *Evolution*, **48**, 1946–58.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–74.

Dillies M-A, Rau A, Aubert J *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, **14**, 671–83.

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

Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–31.

Dubos C, Stracke R, Grotewold E *et al.* (2010) MYB transcription factors in *Arabidopsis*. *Trends in plant science*, **15**, 573–81.

Dyer AG, Whitney HM, Arnold SEJ, Glover BJ, Chittka L (2007) Mutations perturbing petal cell shape and anthocyanin synthesis influence bumblebee perception of *Antirrhinum majus* flower colour. *Arthropod-Plant Interactions*, **1**, 45–55.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–60.

Elleouet J (2012) Evolution of a gene complex controlling flower colour in *Antirrhinum*. MSc Thesis. Université Montpellier II Science et Techniques.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–50.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution*, **64**, 1729–47.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics*, **163**, 939–53.

Feng XZ, Wilson Y, Bowers J *et al.* (2009) Evolution of Allometry in *Antirrhinum*. *Plant Cell*, **21**, 2999–3007.

Fenster CB, Armbruster WS, Wilson P, Dudash MR, Thomson JD (2004) Pollination syndromes and floral specialization. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 375–403.

Ferguson L, Lee SF, Chamberlain N *et al.* (2010) Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Molecular ecology*, **19 Suppl 1**, 240–54.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013) Population genomics from pool sequencing. *Molecular ecology*, **22**, 5561–76.

Flaxman SM, Feder JL, Nosil P (2013) Genetic hitchhiking and the dynamic buildup of genomic divergence during speciationw with gene flow. *Evolution*, **67**, 2577–91.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–18.

Gavrilets S (2010) High-dimensional fitness landscapes and the origins of biodiversity. In: *Toward an Extended Evolutionary Synthesis* (eds Pigliucci M, Muller GB), pp. 45–79. MIT Press.

Gegear RJ, Laverty TM (2005) Flower constancy in bumblebees: a test of the trait variability hypothesis. *Animal Behaviour*, **69**, 939–49.

Geissman TA, Jorgensen EC, Johnson BL (1954) The chemistry of flower pigmentation in *Antirrhinum majus*. Color genotypes. I. The flavonoid components of the homozygous P, M, Y color types. *Archives of Biochemistry and Biophysics*, **49**, 368–88.

Glover BJ, Martin C (1998) The role of petal cell shape and pigmentation in pollination success in *Antirrhinum majus*. *Heredity*, **80**, 778–84.

Goff SA, Cone KC, Chandler VL (1992) Functional analysis of the transcriptional activator encoded by the maize B gene: evidence for a direct functional interaction between two classes of regulatory proteins. *Genes & development*, **6**, 864–75.

Green AA (2007) An integrative analysis of petal growth and form. PhD Thesis. University of East Anglia.

Griffiths A, Miller J, Suzuki D, Lewontin R, Gelbart W (1993) Extensions of mendelian analysis. In: *An Introduction to Genetic Analysis* 5th Edition, pp. 87–117. W. H. Freeman and Company.

Hedrick P (2011) *Genetics of Populations*. 4th Edition Jones and Bartlett Publishers.

Hermann K, Klahre U, Moser M *et al.* (2013) Tight genetic linkage of prezygotic barrier loci creates a multifunctional speciation island in *Petunia*. *Current biology*, **20**, 873–7.

Hoballah ME, Gubitz T, Stuurman J *et al.* (2007) Single gene-mediated shift in pollinator attraction in *Petunia*. *Plant Cell*, **19**, 779–90.

Jiggins CD, Estrada C, Rodrigues A (2004) Mimicry and the evolution of premating isolation in *Heliconius melpomene* Linnaeus. *Journal of evolutionary biology*, **17**, 680–91.

Jiggins CD, McMillan WO (1997) The genetic basis of an adaptive radiation: warning colour in two *Heliconius* species. *Proceedings of the Royal Society B: Biological Sciences*, **264**, 1167–75.

Jin H, Cominelli E, Bailey P *et al.* (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. *The EMBO journal*, **19**, 6150–61.

Joron M, Frezal L, Jones RT *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–6.

Joron M, Papa R, Beltrán M *et al.* (2006) A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS Biology*, **4**, e303.

Kevan P, Giurfa M, Chittka L (1996) Why are there so many and so few white flowers? *Trends in Plant Science*, **1**, 252.

Khimoun A, Burrus M, Andalo C *et al.* (2011) Locally asymmetric introgressions between subspecies suggest circular range expansion at the *Antirrhinum majus* global scale. *Journal of Evolutionary Biology*, **24**, 1433–41.

Khimoun A, Cornuault J, Burrus M *et al.* (2012) Ecology predicts parapatric distributions in two closely related *Antirrhinum majus* subspecies. *Evolutionary Ecology*, **27**, 51–64.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011a) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One*, **6**, e15925.

Kofler R, Pandey R V, Schlotterer C (2011b) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–6.

Konieczny A, Ausubel FM (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant Journal*, **4**, 403–10.

Laitinen RAE, Ainasoja M, Broholm SK, Teeri TH, Elomaa P (2008) Identification of target genes for a MYB-type anthocyanin regulator in *Gerbera hybrida*. *Journal of experimental botany*, **59**, 3691–703.

Lexer C, Widmer A (2008) The genic view of plant speciation: recent progress and emerging questions. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3023–36.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Lindtke D, González-Martínez SC, Macaya-Sanz D, Lexer C (2013) Admixture mapping of quantitative traits in *Populus* hybrid zones: power and limitations. *Heredity*, **111**, 474–85.

Linnen CR, Poh Y-P, Peterson BK *et al.* (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science*, **339**, 1312–6.

Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH (2008) The strength and genetic basis of reproductive isolating barriers in flowering plants. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3009–21.

Lowry DB, Sheng CC, Lasky JR, Willis JH (2012) Five anthocyanin polymorphisms are associated with an R2R3-MYB cluster in *Mimulus guttatus* (Phrymaceae). *American journal of botany*, **99**, 82–91.

Lunau K, Maier EJ (1995) Innate colour preferences of flower visitors. *Journal of Comparative Physiology A*, **177**, 1–19.

Lunau K, Wacht S, Chittka L (1996) Colour choices of naive bumble bees and their implications for colour perception. *Journal of Comparative Physiology A*, **178**, 477–89.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, **21**, 936–9.

Lynch M (2010) Evolution of the mutation rate. *Trends in genetics*, **26**, 345–52.

Lytle BL, Song J, de la Cruz NB *et al.* (2009) Structures of two *Arabidopsis thaliana* major latex proteins represent novel helix-grip folds. *Proteins*, **76**, 237–43.

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends in plant science*, **12**, 57–63.

Mallet J, Barton NH (1989) Strong natural selection in a warning-color hybrid zone. *Evolution*, **43**, 421–31.

Mallet J, Barton N, Lamas G *et al.* (1990) Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones. *Genetics*, **124**, 921–36.

Mallet J, Joron M (1999) Evolution Of Diversity In Warning Color And Mimicry: Polymorphisms, Shifting Balance, and Speciation. *Annual Review of Ecology and Systematics*, **30**, 201–33.

Mandák B, Bímová K, Mahelka V, Placková I (2006) How much genetic variation is stored in the seed bank? A study of *Atriplex tatarica* (Chenopodiaceae). *Molecular ecology*, **15**, 2653–63.

Martin NH, Bouck AC, Arnold ML (2007) The genetic architecture of reproductive isolation in Louisiana irises: Flowering phenology. *Genetics*, **175**, 1803–12.

Martin C, Carpenter R, Sommer H, Saedler H, Coen ES (1985) Molecular analysis of instability in flower pigmentation of *Antirrhinum majus*, following isolation of the pallida locus by transposon tagging. *The EMBO journal*, **4**, 1625–30.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome research*, **23**, 1817–28.

Martin A, McCulloch KJ, Patel NH *et al.* (2014) Multiple recent co-options of Optix associated with novel traits in adaptive butterfly wing radiations. *EvoDevo*, **5**, 7.

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

Matsui K, Umemura Y, Ohme-Takagi M (2008) AtMYBL2, a protein with a single MYB domain, acts as a negative regulator of anthocyanin biosynthesis in *Arabidopsis*. *Plant journal*, **55**, 954–67.

Matus JT, Loyola R, Vega A *et al.* (2009) Post-veraison sunlight exposure induces MYB-mediated transcriptional regulation of anthocyanin and flavonol synthesis in berry skins of *Vitis vinifera*. *Journal of experimental botany*, **60**, 853–67.

Merrill RM, Chia A, Nadeau NJ (2014) Divergent warning patterns contribute to assortative mating between incipient *Heliconius* species. *Ecology and evolution*, **4**, 911–7.

Merrill RM, Gompert Z, Dembeck LM *et al.* (2011) Mate preference across the speciation continuum in a clade of mimetic butterflies. *Evolution*, **65**, 1489–500.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution*, **28**, 659–69.

Michalak P (2009) Epigenetic, transposon and small RNA determinants of hybrid dysfunctions. *Heredity*, **102**, 45–50.

Michel AP, Sim S, Powell THQ *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*, **107**, 9724–9.

Mitchell RJ, Flanagan RJ, Brown BJ, Waser NM, Karron JD (2009) New frontiers in competition for pollination. *Annals of botany*, **103**, 1403–13.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**, 621–8.

Myles S, Peiffer J, Brown PJ *et al.* (2009) Association mapping: critical considerations shift from genotyping to experimental design. *The Plant cell*, **21**, 2194–202.

Nadeau NJ, Martin SH, Kozak KM *et al.* (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular ecology*, **22**, 814–26.

Nadeau NJ, Ruiz M, Salazar P *et al.* (2014) Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome research*, **24**, 1316–33.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 343–53.

Nagy ES, Rice KJ (1997) Local adaptation in two subspecies of an annual plant: implications for migration and gene flow. *Evolution*, **51**, 1079–89.

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The *Arabidopsis* TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *The Plant cell*, **13**, 2099–114.

Niovi Jones K, Reithel JS (2001) Pollinator-mediated selection on a flower color polymorphism in experimental populations of Antirrhinum (Scrophulariaceae). *American journal of botany*, **88**, 447–54.

Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–44.

Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 332–42.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular ecology*, **18**, 375–402.

Nosil P, Parchman TL, Feder JL, Gompert Z (2012) Do highly divergent loci reside in genomic regions affecting reproductive isolation? A test using next-generation sequence data in *Timema* stick insects. *BMC evolutionary biology*, **12**, 164.

Orr HA (1996) Dobzhansky, Bateson, and the genetics of speciation. *Genetics*, **144**, 1331–5.

Orr HA (2001) The genetics of species differences. *Trends in Ecology & Evolution*, **16**, 343–50.

Orr HA, Masly JP, Presgraves DC (2004) Speciation genes. *Current opinion in genetics & development*, **14**, 675–9.

Oyama RK, Jones KN, Baum DA (2010) Sympatric sister species of Californian *Antirrhinum* and their transiently specialized pollinators. *The American Midland Naturalist*, **164**, 337–47.

Pardo-Diaz C, Jiggins CD (2014) Neighboring genes shaping a single adaptive mimetic trait. *Evolution & development*, **16**, 3–12.

Pardo-Diaz C, Salazar C, Baxter SW *et al.* (2012) Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genetics*, **8**, e1002752.

Paz-Ares J, Ghosal D, Wienand U, Peterson PA, Saedler H (1987) The regulatory c1 locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *The EMBO journal*, **6**, 3553–8.

Phillips PC (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9**, 855–67.

Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nature Reviews Genetics*, **11**, 175–80.

Quattrocchio F, Wing J, van der Woude K *et al.* (1999) Molecular analysis of the anthocyanin2 gene of *Petunia* and its role in the evolution of flower color. *Plant Cell*, **11**, 1433–44.

Ramsay NA, Glover BJ (2005) MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. *Trends in plant science*, **10**, 63–70.

Rasmussen IR, Brødsgaard B (1992) Gene flow inferred from seed dispersal and pollinator behaviour compared to DNA analysis of restriction site variation in a patchy population of *Lotus corniculatus* L. *Oecologia*, **89**, 277–83.

Reed RD, Papa R, Martin A *et al.* (2011) optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, **333**, 1137–41.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature communications*, **4**, 1827.

Ridley M (2004) *Evolution*. 3rd Edition Blackwell Publishing.

Rieseberg LH, Blackman BK (2010) Speciation genes in plants. *Annals of botany*, **106**, 439–55.

Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–9.

Ruperti B, Bonghi C, Ziliotto F *et al.* (2002) Characterization of a major latex protein (MLP) gene down-regulated by ethylene during peach fruitlet abscission. *Plant Science*, **163**, 265–72.

Saetre G-P, Kral K, Bures S, Ims RA (1999) Dynamics of a clinal hybrid zone and a comparison with island hybrid zones of flycatchers (*Ficedula hypoleuca* and *F. albicollis*). *Journal of Zoology*, **247**, 53–64.

Schenk MF, Cordewener JHG, America AHP *et al.* (2009) Characterization of PR-10 genes from eight *Betula* species and detection of Bet v 1 isoforms in birch pollen. *BMC plant biology*, **9**, 24.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–41.

Schwinn KE (1999) Regulation of anthocyanin biosynthesis in *Antirrhinum majus*. PhD Thesis. University of East Anglia.

Schwinn K, Venail J, Shang YJ *et al.* (2006) A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. *Plant Cell*, **18**, 831–51.

Shang YJ, Venail J, Mackay S *et al.* (2011) The molecular basis for venation patterning of pigmentation and its effect on pollinator attraction in flowers of *Antirrhinum*. *New Phytologist*, **189**, 602–15.

Simpson JT, Wong K, Jackman SD *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome research*, **19**, 1117–23.

Smith J, Kronforst MR (2013) Do *Heliconius* butterfly species exchange mimicry alleles? *Biology letters*, **9**, 20130503.

Smithson A, Macnair MR (1996) Frequency-dependent selection by pollinators: mechanisms and consequences with regard to behaviour of bumblebees *Bombus terrestris* (L.) (Hymenoptera: Apidae). *Journal of Evolutionary Biology*, **9**, 571–88.

Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. *Annual Review of Plant Biology*, **60**, 561–88.

Sommer H, Saedler H (1986) Structure of the chalcone synthase gene of *Antirrhinum majus*. *Molecular & General Genetics*, **202**, 429–34.

Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. (MA. Noor, Ed,). *PLoS Biology*, **5**, e219.

Stracke R, Werber M, Weisshaar B (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. *Current Opinion in Plant Biology*, **4**, 447–56.

Strasburg JL, Sherman NA, Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 364–73.

Strauss S, Whitall J (2006) Non-pollinator agents of selection on floral traits. In: *Ecology and Evolution of Flowers* (eds Harder L, Barrett S), pp. 120–38. Oxford University Press.

Streisfeld MA, Young WN, Sobel JM (2013) Divergent selection drives genetic differentiation in an R2R3-MYB transcription factor that contributes to incipient speciation in *Mimulus aurantiacus*. *PLoS Genetics*, **9**, e1003385.

Stubbe H (1966) *Genetik und zytologie von Antirrhinum l. Sect. Antirrhinum*. Jena: Gustav Fischer.

Suchet C, Dormont L, Schatz B *et al.* (2010) Floral scent variation in two *Antirrhinum majus* subspecies influences the choice of naïve bumblebees. *Behavioral Ecology and Sociobiology*, **65**, 1015–27.

Supple MA, Hines HM, Dasmahapatra KK *et al.* (2013) Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome research*, **23**, 1248–57.

Tastard E, Andalo C, Giurfa M, Burrus M, Thébaud C (2008) Flower colour variation across a hybrid zone in *Antirrhinum* as perceived by bumblebee pollinators. *Arthropod-Plant Interactions*, **2**, 237–46.

The Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–8.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–11.

Turner TL, Hahn MW (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, **19**, 848–50.

Turner TL, Hahn MW, Nuzhdin S V (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.

Turner ME, Stephens JC, Anderson WW (1982) Homozygosity and patch structure in plant populations as a result of nearest-neighbor pollination. *Proceedings of the National Academy of Sciences*, **79**, 203–7.

Vargas P, Carrio E, Guzman B, Amat E, Guemes J (2009) A geographical pattern of *Antirrhinum* (Scrophulariaceae) speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. *Journal of Biogeography*, **36**, 1297–312.

Vargas P, Ornosa C, Ortiz-Sánchez FJ, Arroyo J (2010) Is the occluded corolla of *Antirrhinum* bee-specialized? *Journal of Natural History*, **44**, 1427–43.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular ecology*, **14**, 3623–42.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 451–60.

Wallace B (1991) Coadaptation revisited. *The Journal of heredity*, **82**, 89–95.

Wang L, Luzynski K, Pool JE *et al.* (2011) Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. *Molecular ecology*, **20**, 2985–3000.

Whibley AC (2004) Molecular and genetic variation underlying the evolution of flower colour in *Antirrhinum*. PhD Thesis. University of East Anglia.

Whibley AC, Langlade NB, Andalo C *et al.* (2006) Evolutionary paths underlying flower color variation in *Antirrhinum*. *Science*, **313**, 963–6.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular ecology*, **19**, 925–39.

Whitney HM, Glover BJ (2007) Morphology and development of floral features recognised by pollinators. *Arthropod-Plant Interactions*, **1**, 147–58.

Whitney HM, Milne G, Rands SA *et al.* (2013) The influence of pigmentation patterning on bumblebee foraging from flowers of *Antirrhinum majus*. *Die Naturwissenschaften*, **100**, 249–56.

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. *PloS One*, **7**, e42649.

Wilson Y, Hudson A (2011) The evolutionary history of *Antirrhinum* suggests that ancestral phenotype combinations survived repeated hybridizations. *Plant Journal*, **66**, 1032–43.

Winkel-Shirley B (2001) Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant physiology*, **126**, 485–93.

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, **8**, 206–16.

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, **1**, 356 – 66.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–65.

Xue Y, Carpenter R, Dickinson HG, Coen ES (1996) Origin of allelic diversity in *Antirrhinum* S locus RNases. *The Plant cell*, **8**, 805–14.

Yuan Y-W, Sagawa JM, Young RC, Christensen BJ, Bradshaw HD (2013) Genetic dissection of a major anthocyanin QTL contributing to pollinator-mediated reproductive isolation between sister species of *Mimulus*. *Genetics*, **194**, 255–63.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–29.

Zhang J (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, **18**, 292–98.

Zheng W, Chung LM, Zhao H (2011) Bias detection and correction in RNA-Sequencing data. *BMC bioinformatics*, **12**, 290.