Evolutionary genetics of flower colour variation in *Antirrhinum*

Mabon Rhun Elis

Thesis submitted
for the Degree of
Doctor of Philosophy

University of East Anglia
John Innes Centre

Submitted in March 2018

ABSTRACT

Phenotypic differences between species and populations can reveal much about how they have adapted and responded to a complex set of environmental cues. Studies have shown that genetic control of some traits is centralised to single genomic regions, while others are regulated at many unlinked loci dispersed throughout the genome. One trait that shows an enormous degree of variation between plant species is flower colour, and its tractability makes it an ideal trait for studying genetic differences underlying species differentiation. *Antirrhinum majus* has long been used as a model for studying floral traits, including colour. The 20-30 wild *Antirrhinum* species use diverse patterns on their flowers, formed by producing and accumulating magenta anthocyanins and yellow aurones in different tissues, to attract pollinators. In this project, I sought to genetically map flower colour phenotypes to the *Antirrhinum* genome. Several *Antirrhinum* species were crossed to *A. majus* to generate segregating populations. I used a combination of bulked segregant analysis, individual genotyping of segregating populations and analysis of genome sequences from wild accessions to test whether genes governing each colour trait were concentrated at particular loci or dispersed across many chromosomes. I found that variation in magenta not previously characterised maps to the known *ROSEA-ELUTA* (*ROS-EL*) locus where transcription factors regulating anthocyanin production are encoded. Yellow phenotypes from three species mapped to chromosome 2, where there is reduced recombination between *A. majus* and many wild species, and where an aurone biosynthetic enzyme is encoded. However, there appear to be some additional modifiers of flower colour in these species, not linked to the *ROS-EL* and chromosome 2 loci. These results fit neither the central- nor dispersed-control models of genetic control, but rather an intermediate hypothesis – that flower colour can be changed by selection acting on a modest number of loci spread throughout the genome.

CRYNODEB

Gall gwahaniaethau mewn ffenoteip rhwng rhywogaethau a phoblogaethau ddatgelu llawer am y modd maen nhw wedi addasu drwy ymateb i gyfresi cymhleth o arwyddion amgylcheddol. Mae ymchwil wedi dangos fod rheolaeth enetig rhai nodweddion wedi ei ganoli ar rannau penodol o'r genom, tra bo rheolaeth nodweddion eraill yn digwydd mewn nifer o rannau di-gyswllt wedi eu gwasgaru ar hyd y genom. Un nodwedd sy'n arddangos amrywiaeth aruthrol rhwng rhywogaethau o blanhigion yw lliw blodau. Mae hydrinedd y nodwedd hon yn ei gwneud yn un ddelfrydol ar gyfer astudio newidiadau genetig sy'n tanseilio gwahaniaethau rhwng rhywogaethau. Mae'r planhigyn *Antirrhinum majus* wedi ei ddefnyddio ar gyfer astudio nodweddion mewn blodau, gan gynnwys lliw, ers dros ganrif. Mewn rhywogaethau cysylltiedig – y 20-30 aelod o'r genws *Antirrhinum* a geir yn y Canoldir – gwelir patrymau amryfath ar flodau'r planhigion. Ffurfir y lliwiau a'r patrymau yma gan gynhyrchu pigmentau majenta (anthocyanin) a melyn (aurone) mewn rhannau gwahanol o'r petalau, er mwyn denu gwenyn fel peillwyr. Yn yr ymchwil yma, edrychais ar ffenoteipiau lliw rhai o'r gwahanol rywogaethau a'u mapio i enom *Antirrhinum*. Fe groeswyd llawer o rywogaethau gyda *A. majus* i greu poblogaethau yn arwahanu am ffenoteipiau gwahanol. Defnyddiais nifer o ddulliau genetig a genomig i brofi dwy ddamcaniaeth gyferbyniol – ydi ffenoteipiau lliw blodau wedi eu rheoli yn ganolog ar un locws neu ar wasgar drwy'r genom? Fe welais fod amrywiaeth mewn patrwm majenta oedd heb ei hynodi o'r blaen wedi ei reoli ar locws oedd wedi ei ddisgrifio yn barod, a bod llawer o ffenoteipiau melyn gwahanol yn mapio i'r un mannau a'i gilydd. Fodd bynnag, gwelais hefyd fod gan rai o'r ffenoteipiau yma addaswyr gwahanol wedi eu lleoli mewn rhannau eraill o'r genom. Dangosa hyn fod rheolaeth enetig yn ffitio damcaniaeth ganolraddol i'r ddwy a osodais ar ddechrau'r prosiect.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND NOMENCLATURE

ANS	anthocyanidin synthase
AS1	*AUREUSIDIN SYNTHASE1*, which encodes AUS
AUS	aureusidin synthase
BLAST	Basic Local Alignment Search Tool
BSA	bulked segregant analysis
BWA	Burrows-Wheeler Aligner
CGT	chalcone glucosyltransferase
CHI	chalcone isomerase
CHS	chalcone synthase
CTAB	cetrimonium bromide
DEPC	diethyl pyrocarbonate
DFR	dihydroflavonol reductase
EDTA	ethylenediaminetetraacetic acid
EL	*ELUTA*
F3'5'H	flavonoid 3'5'-hydroxylase
F3H	flavanone 3-hydroxylase
F3'H	flavonoid 3'-hydroxylase
FAM	fluorophore giving fluorescent signal in the blue channel
FLA	FLAVIA, which encodes CGT
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GATK	Genome Analysis Toolkit
GWAS	genome-wide association study
JIC	John Innes Centre
KASP	kompetitive allele-specific PCR
LAR1	*LIGHT AREAS1* (*Mimulus lewisii*)
NIL	near-isogenic line
PCR	polymerase chain reaction
QTL	quantitative trait locus
Qubit	method for fluorometric quantitation of nucleic acids
RFU	relative frequency unit
ROS	*ROSEA*
SNP	single nucleotide polymorphism
SULF	*SULFUREA*
TGAC	The Genome Analysis Centre (now the Earlham Institute)
UF3GT	UDP-glucose:flavonoid 3-o-glucosyltransferase
UF3RT	UDP-glucose:flavonoid 3-o-rhamnosyltransferase
VE	*VENOSA*
VIC	fluorophore giving fluorescent signal in the green channel
WGS	whole genome sequencing

ACKNOWLEDGMENTS

My thanks go to:

Enrico Coen, my supervisor, for his continued support throughout this project and for the ideas that drove this research.

Desmond Bradley for his help with everything from lab work and plant work to fantastically helpful comments on this thesis and my research ideas.

Annabel Whibley for her help and advice on so many issues from bioinformatics to evolution.

Lucy Copsey, without whom this would be a thesis about failing to grow and cross *Antirrhinum*.

Matt Couchman for all his bioinformatics help.

Hugo Tavares and Louis Boell for their help during the early years of this project.

Hybrid zone collaborators, including Christophe Andalo, Monique Burrus, David Field, Nick Barton and countless students and volunteers.

My secondary supervisor, Levi Yant, for his advice and ideas.

Yongbiao Xue for providing access to the unpublished *Antirrhinum* genome and for most of the sequencing in this thesis.

Everyone in the Coen Lab I've had the pleasure of working with for four years.

All the friends I've made during my PhD, who have supported me and shared this journey with me.

Diolch i fy nheulu – Dad, Mam, Cai a Siriol – am eu cariad ac am fy nghefnogi dros y blynyddoedd.

And thanks to Alex for his love, support and ability to spot many silly mistakes in this thesis.

FUNDING

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

1 Introduction

1.1 Adaptive variation and its genetic regulation

Organisms adapt to a complex set of environmental cues. Within a species, this adaptive response results in trait variation that is maintained within and between populations through a balance of mutation, genetic drift and natural selection. When such polymorphisms go to fixation, this results in species-level trait differences (Charlesworth and Charlesworth 2010). This can be seen in the diversity found in the natural world.

Evolutionary biologists have long been interested in studying trait variation in the wild and the genetic variation that underlies it. During his voyages on the Beagle, Charles Darwin observed the variation between inhabitants of different islands in the Galapagos. Here, organisms showed specialised adaptations to small-scale habitat differences. This is best-illustrated in his famous finches (*Geospiza* spp), in which natural selection has led to different beak shapes and sizes according to food availability and the birds' feeding preferences (Darwin 1859). This has established the groundwork for more than 150 years of observational studies and genetic analyses of traits under selection. The field now includes work on thousands of different plant and animal species, all of which show a diversity of traits as adaptations to their

environments.

1.2 Colour variation

Nowhere is trait variation more striking than in the diversity of colours and patterns seen in nature. Organisms produce colours as warning signals to predators, camouflage to hide in different environments, patterns to attract mates or signal to pollinators, and to function in metabolic processes such as photosynthesis. Some additional colour variation is of unknown evolutionary advantage or may be selectively neutral.

Flowering plants (angiosperms) are a monotypic clade of 369,000 species (Willis 2017) that have recruited colour in various organs for different functions. In flowers, colours and patterns are finely-tuned to manipulate animals in a way that maximises the plants' reproductive success. Colour is used, alongside other cues such as scent, to attract pollinators from a distance, guide them towards specific parts of the flower and, sometimes, in specialised ways such as mimicry of insect-rewarding flowers by nectarless species (Schiestl and Johnson 2013).

Many of these colours and patterns have evolved as a result of selection effected by animals that interact with the plants, giving rise to convergent evolution of similar colours in plants pollinated by the same animals (pollination syndromes), as well as highly-specialised plant-pollinator pairings involving mimicry. Red flowers are typically pollinated by birds, which have good red colour vision unlike most insects. Bee-pollinated flowers tend to be yellow or purple, colours that are better-detected by these insects' visual receptors (Schiestl and Johnson 2013). Contrasting flower colours can often be seen in closely-related species that attract different pollinators. The scarlet monkeyflower, *Mimulus cardinalis* (Phrymaceae), has bright red flowers to attract hummingbirds, while its close relative, the great purple monkeyflower, *M. lewisii*, has purple-magenta flowers to attract bees. Experimental evidence has shown that flower colour alone can lead to a change in pollinator from bumblebees to hummingbirds in these *Mimulus* species (Bradshaw and Schemske 2003).

M. lewisii also uses patches of yellow pigment surrounded by regions of white to guide bees towards a landing patch (Owen and Bradshaw 2011, Yuan *et al* 2016). Such 'nectar guides' are important in plants as they help increase pollinator efficiency and,

thus, pollination efficiency (Hansen *et al* 2012). These guides range in complexity from contrasting spots in particular flower regions (**Figure 1.1 a**) to a combination of colours that produce more elaborate patterns (**Figure 1.1 b**). Some plants also use structural colour to highlight regions on their flowers to pollinators (**Figure 1.1 c**). This produces angle-dependent colour patterns (iridescence) that have been shown to increase pollinator efficiency (Moyroud *et al* 2017). Plants can also use a combination of different pigment colours, structural colours, textures and chemical signals to sexually mimic pollinators, resulting in a highly specialised plant-pollinator relationship where the plants are pollinated by insects attempting copulation (Devey *et al* 2008) (**Figure 1.1 d**).

Figure 1.1 Flower traits used by plants to attract and guide pollinators: a *Mimulus guttatus* flower with red spots on its landing patch (**a**); an *Impatiens* flower with magenta and yellow nectar guides that converge around the centre of the flower (photograph by Alex Twyford) (**b**); a *Hibiscus trionum* flower with a dark iridescent ring, from Vignolini *et al* (2014) (**c**); and an *Ophrys* flower with an elaborate labellum that resembles the insects that pollinate the plants (**d**).

1.3 Biochemistry of flower colour

1.3.1 Flower colour pigments

Plants produce pigments using several different pathways. Many of these pigments are used in flowers to attract pollinators, although some accumulate in other tissues and contribute towards different plant processes. The pigments used in flowers fall into one of three classes of compounds: flavonoids, betalains and carotenoids.

1.3.1.1 Flavonoids

Flavonoids are secondary metabolites that serve a diverse range of functions in land

plants. The complex pathway that produces them, discussed in detail in section 1.5.2 on page 28, has evolved gradually in plants, with more recently evolved clades producing novel classes of flavonoids compared to ancient ones (**Figure 1.2**) (Rausher 2006). Although famed for the colours they confer to flowers and other plant organs, flavonoids play a wide range of roles in plant physiology, including deterring herbivores, protecting tissues against damage from ultraviolet light and oxidation, and mediating symbioses between plants and fungi (Koes *et al* 1994).

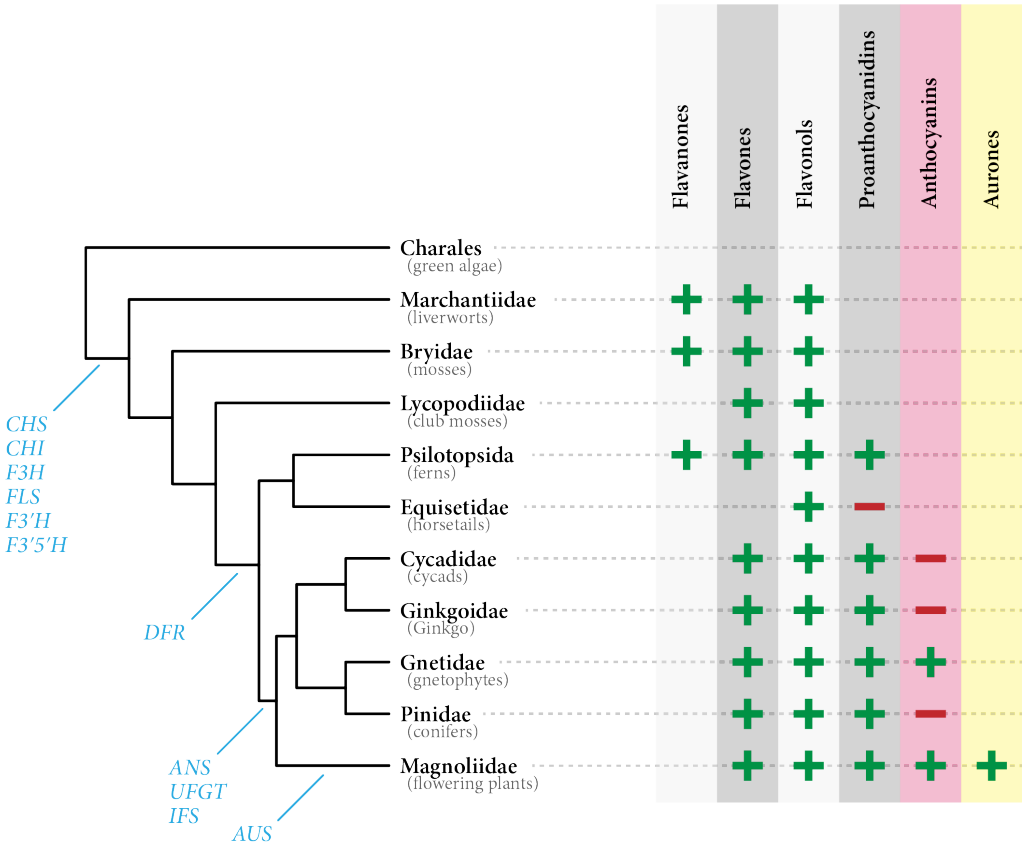


Figure 1.2 Phylogeny of land plants based on Chase and Reveal (2009) annotated with the time of origin of flavonoid biosynthetic enzymes (blue labels) and the occurrence of six classes of flavonoids, adapted from Rausher (2006). A green plus (+) indicates the documented presence of a flavonoid class; a red minus (−) indicates the possible evolutionary loss of a flavonoid class. Enzyme abbreviations: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; FLS, flavonol synthase; F3′H, flavonoid 3′-hydroxylase; F3′5′H, flavonoid 3′5′-hydroxylase DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; UFGT, UDP-glucose: flavonoid glucosyltransferase; IFS, isoflavone synthase; AUS, aureusidin synthase.

Several flavonoids are involved in flower colour. The best-studied examples are

anthocyanins, which give flowers red to blue hues, depending on the degree to which the B ring (**Figure 1.3 a**) of the molecular backbone is hydroxylated and/or O-methylated (**Figure 1.3 b**). The three major anthocyanins are the 3-glucosides of pelargonidin, cyanidin and delphinidin, which give orange-red, magenta and purple colours, respectively (Glover and Martin 2012).

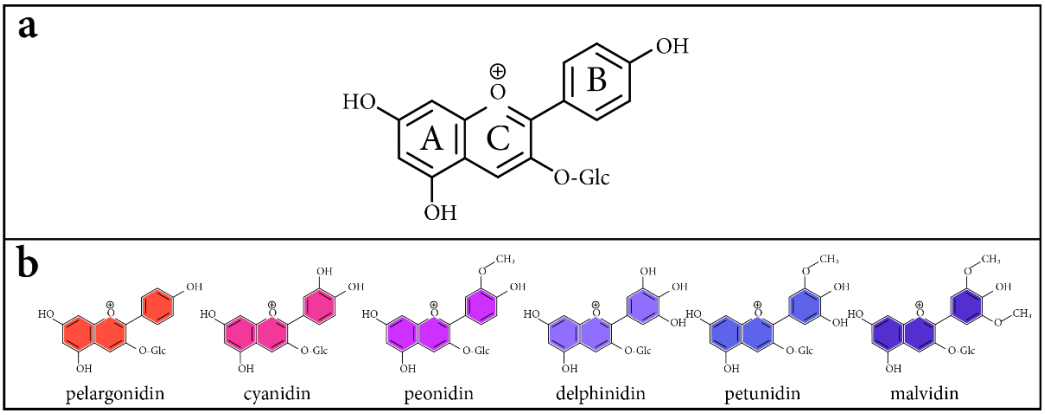


Figure 1.3 An anthocyanin molecule showing the A, B and C rings, adapted from Glover and Martin (2012) (**a**); the red to blue hues given by anthocyanins according to the degree to which the B ring is hydroxylated, adapted from Ananga *et al* (2013) (**b**).

Another class of flavonoids important for flower colour is aurones. These pigments give a bright yellow colour to several flowering plant species, mostly in the Plantaginaceae and Asteraceae families (Nakayama 2002). They are considered to have evolved later than most other major flavonoids and are only found in the flowering plants (Rausher 2006). Chalcone, a precursor of flavonoids, can also give a pale-yellow colour to flowers, including in some *Dianthus* and *Cyclamen* species.

1.3.1.2 Betalains

Some plants with red- and purple-coloured organs in the order Caryophyllales use a separate class of pigments, following the loss of anthocyanins from most members of that order. These betalain pigments give *Bougainvillea* bracts, *Nepenthes* pitcher plant traps and Christmas cactus (*Schlumbergera*) flowers their characteristic colours (Strack *et al* 2003).

1.3.1.3 Carotenoids

Most yellow flowers are coloured by carotenoids, a class of lipid-soluble isoprenoid-derived compounds synthesised within plastids (Glover 2014). Carotenoids are some

of the most abundant naturally occurring pigments on earth, produced not only in land plants but also in algae, and in some fungi and bacteria (Nisar *et al* 2015). *Narcissus*, *Brassica* and *Mimulus* flowers are all coloured yellow by carotenoids (Valadon and Mummery 1968, Yuan *et al* 2014, Zhang *et al* 2015).

1.4 Finding genes for phenotypes

Understanding the genetic architecture of phenotypic variation has been an important aim in biology since Sturtevant (1913) first used studies of linkage and recombination to genetically map traits in *Drosophila* fruit flies. The discovery of DNA as the genetic material led to the development of genetic markers, first by looking at enzymes encoded by alternate alleles (allozyme variants) in *Drosophila pseudoobscura* (Hubby and Lewontin 1966), and then by using restriction enzymes to examine polymorphisms in DNA itself (Saiki *et al* 1985).

Developments such as the invention of DNA amplification through polymerase chain reaction (PCR) and the ability to clone and sequence DNA brought with them the desire to understand the relationship between phenotypic differences and variation at the molecular level (Altshuler *et al* 2008). In the late 1980s, genes underlying human diseases were sequenced for the first time (Kerem *et al* 1989), and quantitative trait loci (QTLs) underlying complex continuous traits were identified (Edwards *et al* 1987). Single nucleotide polymorphisms (SNPs) – changes in DNA from one nucleotide base to another – between individuals of the same species were first described by Kreitman (1983). He showed that most SNPs within genes occur without changing the amino acid sequences of the proteins they encode, suggesting that natural selection constrains protein sequences. Advances in high-throughput sequencing technologies in recent years have made discovering SNPs in virtually any natural system possible (Dalziel *et al* 2009). Techniques such as restriction site-associated DNA (RAD) sequencing and whole genome sequencing have made it possible to develop genome-wide markers for genetic variability, useful not only in functional biology, but in evolutionary biology too. By identifying the genetic loci that underlie traits distinguishing species and populations, we can look at the evolutionary forces that shape speciation and divergence (Anderson *et al* 2011).

The ability to study molecular differences in natural populations has massively increased our knowledge of how the genetic control of adaptive traits is organised in the genome. An important question is the relative contributions of different loci to a phenotype: is a trait regulated by one or a few loci of large effect, or do many separate loci dispersed throughout the genome each have marginal contributions to an overall perceived phenotype (King and Long 2017)?

Genome-wide association studies (GWAS) have shown that some traits that vary between populations, and between individuals within a population, are governed by many independent loci. At least 71 loci contribute towards susceptibility to Crohn's disease in humans (Franke *et al* 2010), for example, while many loci involved in such complex traits are likely to go undetected – the so-called Beavis effect (King and Long 2017). At the other extreme, horn polymorphism in Soay sheep (*Ovis aries*) maps to a single locus owing to a balance between sexual and natural selection (Johnston *et al* 2013).

Another example of a single locus associated with multiple phenotypes in natural populations is seen in *Heliconius* butterflies, a genus of mimetic insects found in the neotropics. In several *Heliconius* species, a locus named *OPTIX* encodes a transcription factor that controls red wing patterning (Jiggins *et al* 2017). GWAS results show that differences in red patterning across multiple species (**Figure 1.4**) map to the *OPTIX* locus. Wallbank *et al* (2016) showed that two 50-100 kb sequences in the *cis*-regulatory region of *OPTIX* regulate where the gene is expressed, and that evolutionary shuffling between the two sequences through hybridisation has resulted in different patterns across species. This is thought to allow mimicry of different butterfly species without compromising *OPTIX* function through coding-sequence changes.

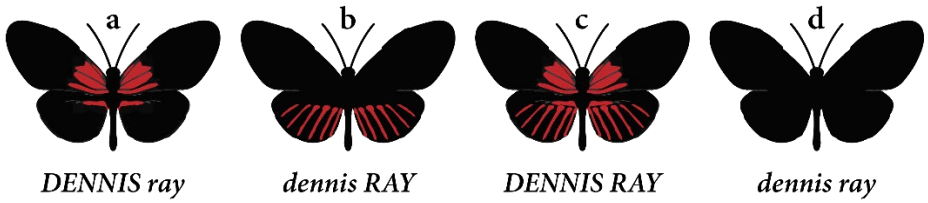


Figure 1.4 Schematic representation of wing patterns in *Heliconius* butterflies caused by two *cis*-regulatory changes, *DENNIS* and *RAY*, in the promoter region of the *OPTIX* gene. Some species, such as *H. meriana*, carry the *DENNIS* sequence in this region, resulting in expression of *OPTIX* and red banding on the upper half of the wings (**a**). Others, such as *H. contigua*, carry the *RAY* sequence, resulting in red bands on the lower half of the wings (**b**). Species carrying both sequences, such as *H. elevatus*, have both patterns (**c**), while those with neither sequence, such as *H. rosina*, do not show either pattern. Patterns regulated by other loci have been removed from the images. *DENNIS* and *RAY* haplotypes are shown below the diagrams. Adapted from Wallbank *et al* (2016).

Genetic changes affecting regulation of gene expression are also crucial in the evolution of flower colour (Streisfeld and Rausher 2011). Many flower colour pigments, including flavonoids and carotenoids, are important not only for reproduction, but for other physiological processes in plants, too. Carotenoids expand the range of wavelengths plants can utilise in photosynthesis (Hashimoto *et al* 2016), and flavonoids are involved in defending plants against a host of biotic and abiotic stresses (Buer *et al* 2010). Thus, it has been proposed that adaptations involving flower colour evolve by differences in the expression of existing genes, with natural selection likely acting against mutations that would have pleiotropic effects on other plant functions (Streisfeld and Rausher 2011). Such differences typically involve mutations in the *cis*-regulatory regions of structural genes involved in pigment production, or in coding or *cis*-regulatory regions of the transcription factors that interact with the structural genes (Wu *et al* 2013). Research on model organisms such as *Mimulus*, *Petunia*, *Ipomoea* and *Antirrhinum* has contributed greatly to our understanding of the different methods plants have evolved for regulating expression of their flower colour genes (Sobel and Streisfeld 2013).

1.5 *Antirrhinum* species and their colours

Antirrhinum majus L. (Plantaginaceae) has been used for over a century as a model organism for studying floral trait variation (Schwarz-Sommer *et al* 2003). Darwin

(1868) was fascinated by the morphology of *Antirrhinum* flowers and the 'peloric' mutants he saw that had radially-symmetrical flowers as opposed to the bilateral symmetry of wildtype flowers. Research in *Antirrhinum* has since made fundamental contributions to research on flower development (Coen and Meyerowitz 1991) and the biosynthesis of flower colour pigments (Martin *et al* 1991), as reviewed in Schwarz-Sommer *et al* (2003) and Hudson *et al* (2008).

The *Antirrhinum* corolla is made up of five petals which are fused for part of their length, forming a tube enclosed by two upper and three lower lobes (**Figure 1.5**). The flowers are pollinated by bees, and their shape is thought to be an adaptation to these pollinators (Vargas *et al* 2010). The closed flower structure requires bees to land on a platform, prise apart the lobes to access nectar at the base of the tube, and at the same time contact the anthers and stigma to pollinate the flower (**Figure 1.6**). This mechanism excludes smaller, lighter insects, and typically only large bees – mostly *Bombus lucorum*, *B. hortorum*, *B. lapidarius* and *Xylocopa violacea* – can gain entry (Tastard *et al* 2012).

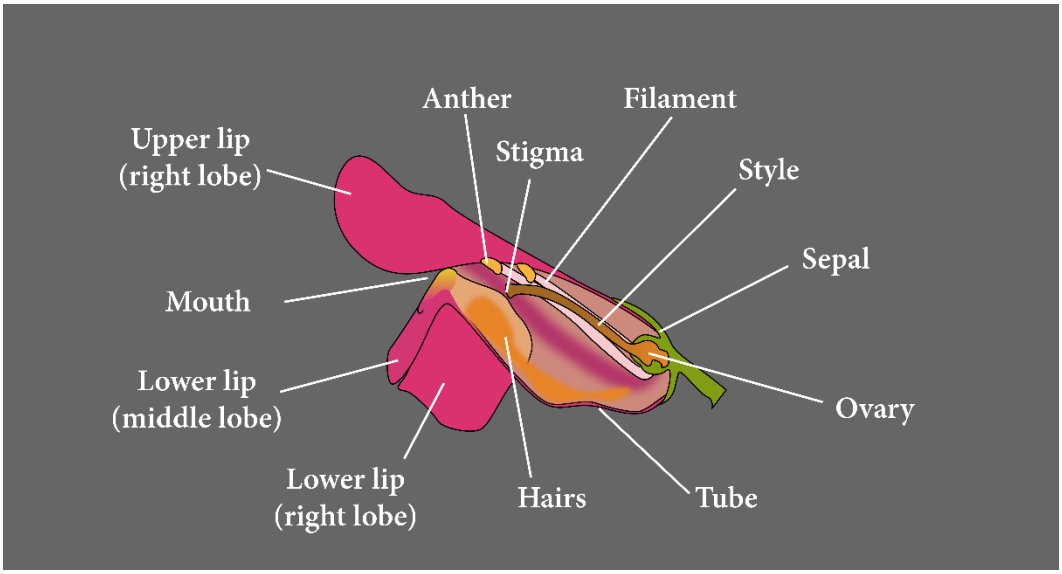


Figure 1.5 Cross-section of an *A. majus* flower, showing the floral organs and their positions.

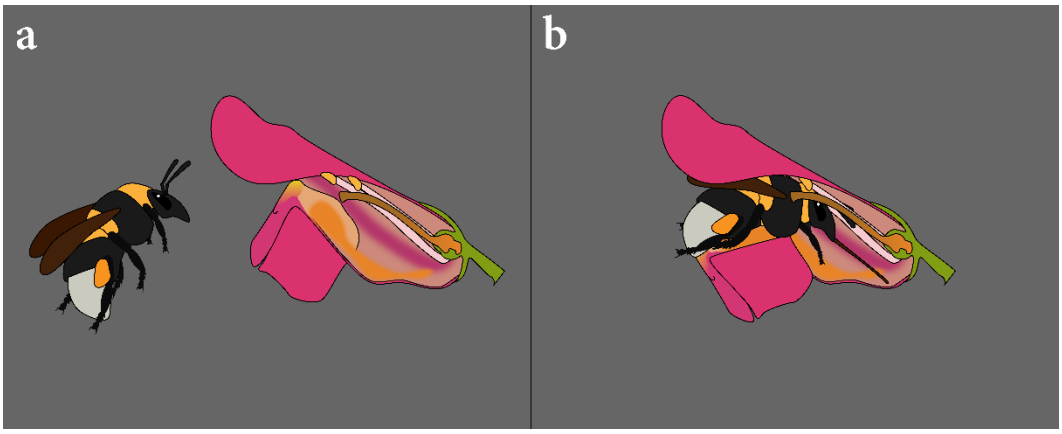


Figure 1.6 Pollination of an *A. majus* flower, with the flower in cross-section. A bumblebee (*B. hortorum*) approaches the flower (**a**). To gain entry, she must land on the lower lobes of the flower, push them down and crawl inside the tube to access the nectar (**b**).

1.5.1 Flower colour in *A. majus*

Two types of flavonoid pigments accumulate in the flowers of *Antirrhinum* species to give them their colours. Magenta colours are produced by anthocyanin – typically cyanidin 3-rutinoside, although cyanidin 3-glucoside has also been found in the flowers of some cultivars (Gilbert 1971). Yellow colours are produced by an aurone named aureusidin glucoside (Nakayama 2002).

1.5.2 The flavonoid biosynthetic pathway

Three enzymes are required to convert a molecule of coumaroyl-CoA and three molecules of malonyl-CoA to aureusidin glucoside (aurone) and eight are required to make cyanidin 3-rutinoside (anthocyanin) (**Figure 1.7**). Many of the structural genes encoding these enzymes were identified in *A. majus*, mostly using mutants with transposable element insertions. The *nivea* mutant has a mutation in the gene encoding chalcone synthase (*CHS*) (Sommer and Saedler 1986), *incolorata* in flavanone 3-hydroxylase (*F3H*) (Martin *et al* 1991), *eosinea* in flavonoid 3′-hydroxylase (*F3′H*) (Stickland and Harrison 1977), *pallida* in dihydroflavonol reductase (*DFR*) (Martin *et al* 1985) and *candica* in anthocyanidin synthase (*ANS*) (Martin *et al* 1991). The structural genes involved in aurone biosynthesis were characterised more recently – *AUREUSIDIN SYNTHASE 1* (AS1), which encodes aureusidin synthase (AUS) by Nakayama *et al* (2000) and the sequence of *FLAVIA* (*FLA*), the gene encoding chalcone

glucosyltransferase, which adds a glucose group to the A ring of chalcone ahead of conversion to aureusidin (Ono *et al* 2006), is not yet published (Boell *et al* unpublished results).

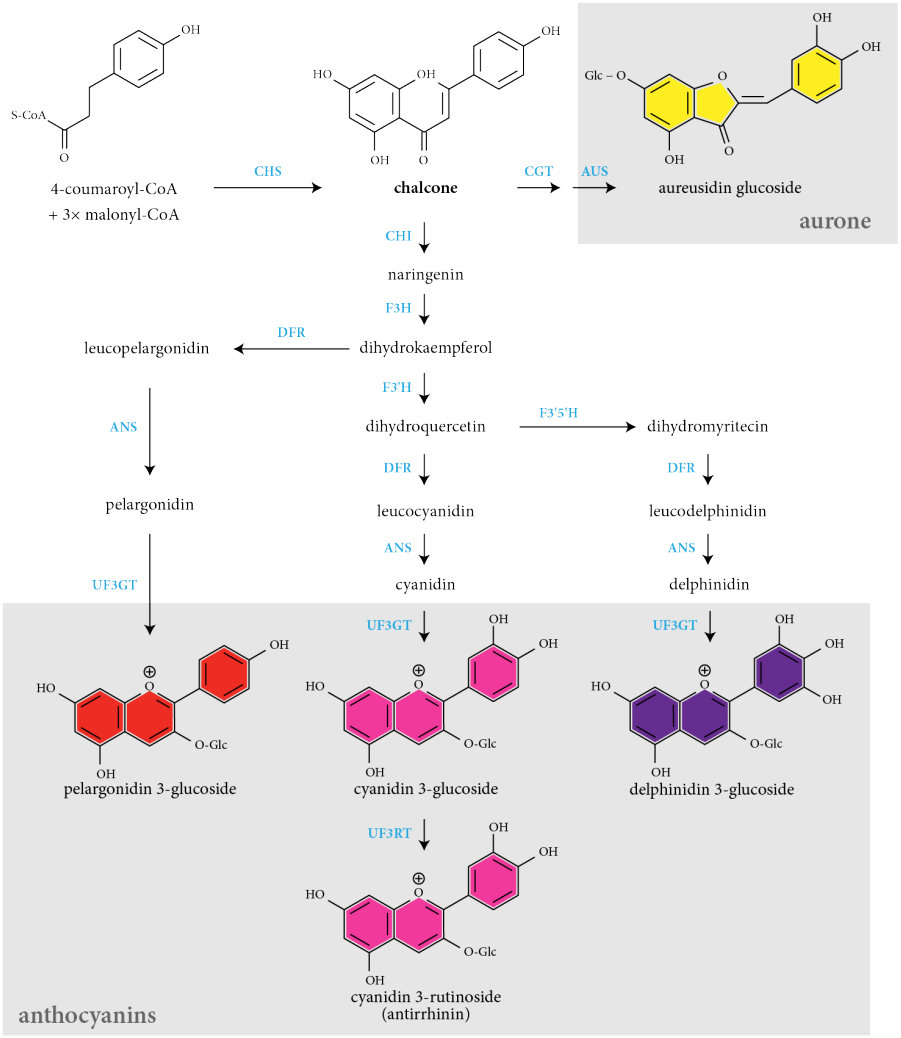


Figure 1.7 Biosynthetic pathway of flavonoids with a focus on the pigments aurones and anthocyanins, which have been coloured according to their appearance in aqueous solutions. Flavonoid names are shown in black, with enzyme names in blue. Enzyme abbreviations: CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3′H, flavonoid 3′-hydroxylase; DFR, dihydroflavonol 4-reductase; F3′5′H, flavonoid 3′5′-hydroxylase; ANS, anthocyanidin synthase; UF3GT, UDP-glucose: flavonoid 3-glucosyltransferase; UF3RT, UDP-glucose: flavonoid 3-rhamnosyltransferase; CGT, chalcone glucosyltransferase; AUS, aureusidin synthase. Adapted from Martin *et al* (1991) and Rausher (2006), with the addition of CGT as characterised by Ono *et al* (2006).

In addition to the structural genes, many genes encoding transcription factors are also involved in regulating flavonoid biosynthesis. These activate structural genes acting late in the pathway (from *F3H* onwards): *DELILA* (*DEL*) encodes a basic helix-loop-helix transcription factor to activate pigmentation in the corolla tube (Martin *et al* 1991); *ROSEA* encodes two MYB-like transcription factors activating these genes throughout the petals (Schwinn *et al* 2006); *VENOSA* encodes another MYB-like transcription factor activating the pathway in tissue overlying veins in the dorsal petals of the flowers; and *ELUTA*, which suppresses the biosynthesis of anthocyanins in parts of the petals (Martin *et al* 1991) was recently shown to be another *MYB* gene (Tavares *et al* in review). These regulatory genes have been shown to underlie variation in magenta pigmentation in wild *Antirrhinum* species (Schwinn *et al* 2006, Whibley *et al* 2006, Shang *et al* 2011, Tavares *et al* in review).

1.5.3 Diversity of *Antirrhinum* species

Between 20 and 30 *Antirrhinum* species, depending on the taxonomic treatment used, are native to the Mediterranean (Rothmaler 1956, Vargas *et al* 2009, Wilson and Hudson 2011). Most of the diversity in the *Antirrhinum* genus is found on the Iberian peninsula, where the majority of species are found (**Figure 1.8**). They occupy diverse habitats and show an extensive amount of variation in organ morphology, growth habit and flower colour (Vargas *et al* 2009, Wilson and Hudson 2011).

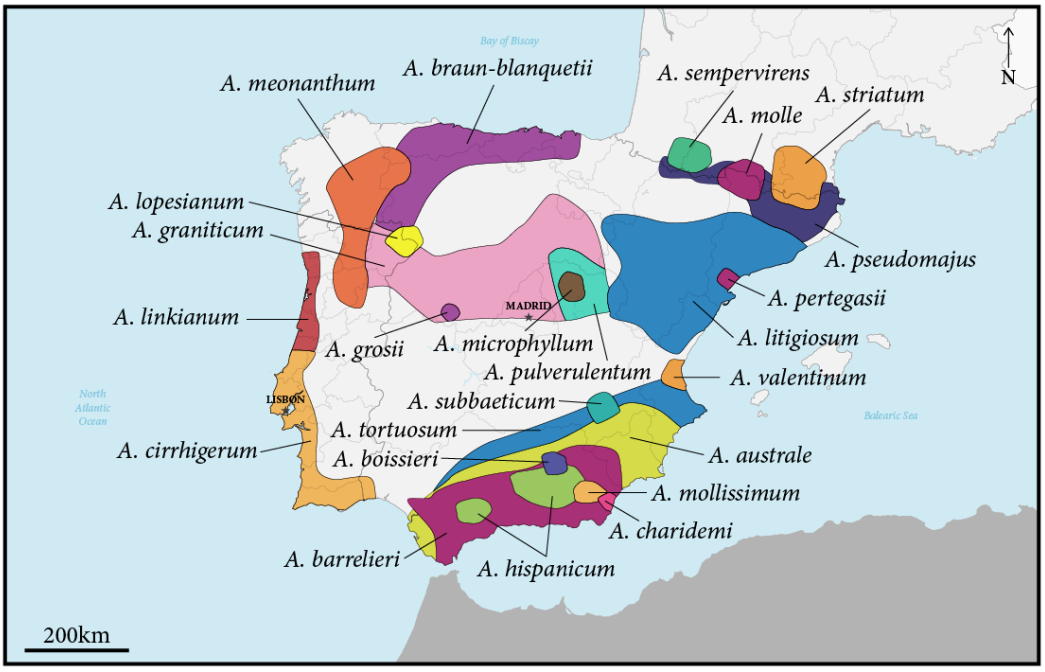


Figure 1.8 Map of *Antirrhinum* species distribution on the Iberian peninsula. Distribution information is taken from Rothmaler (1956), Whibley (2004) and Wilson and Hudson (2011).

Despite the phenotypic diversity across the genus, however, nearly all *Antirrhinum* species are inter-fertile and are thought to have radiated relatively recently – in the last 3-5 million years (Whibley 2004). This recent radiation, coupled with likely hybridisation between species, has made resolving phylogenetic relationships and determining ancestral phenotypes in the genus difficult (Wilson and Hudson 2011).

Antirrhinum species have one of two contrasting growth habits. Twelve of the species whose population distributions are shown in **Figure 1.8** are small, prostrate plants that grow on rocky cliffs. These have small flowers, small, often succulent, leaves, and show good drought and cold tolerance, but are thought to be poor at competing for resources with other plants (**Figure 1.9 a-c**). Others are much larger and grow upright on disturbed (ruderal) habitats. These tend to have large flowers and large, thin leaves (**Figure 1.9 d-f**) (Wilson and Hudson 2011). Flower colour also appears to correlate with these ecological differences. The prostrate, cliff-dwelling species mostly have white or pale pink flowers, with a small amount of colour thought to act as pollinator guides. These guides are seen in the ruderal species too, but are complemented by bright yellow or magenta pigmentation accumulating throughout the petals.

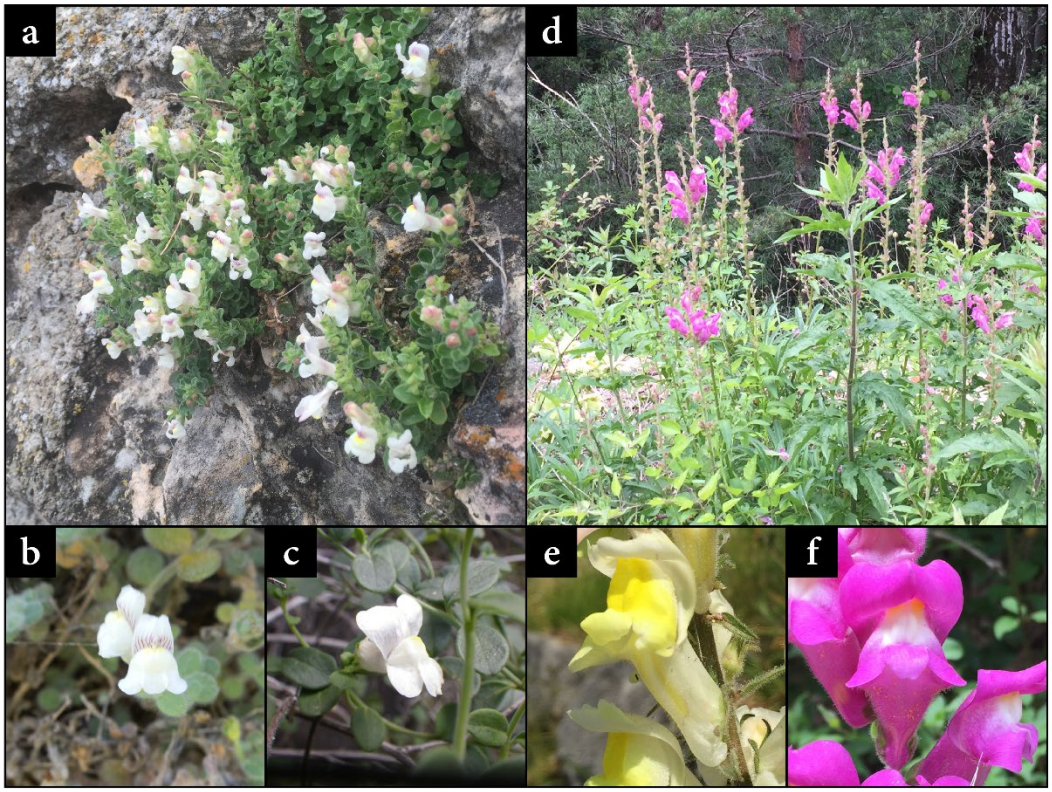


Figure 1.9 Growth habits and flower colours in *Antirrhinum*. Many species, such as *A. pulverulentum* (**a**), *A. lopesianum* (**b**) and *A. valentinum* (**c**), are small, prostrate xerophytes growing in isolation on cliff faces. These usually have small white or pale pink flowers. Others, such as *A. pseudomajus* (**d** and **f**) and *A. braun-blanquetii* (**e**), are large and upright and are found growing in competition with other plants. These have large, brightly-coloured flowers. Photographs c, d and f taken by Enrico Coen.

The colours and patterns seen on *Antirrhinum* flowers are adaptations to help the plants attract and guide bees. The bright colours are believed to attract the insects from a distance (Whibley *et al* 2006, Bradley *et al* 2017), while markers such as magenta veins on the upper lobes and yellow 'foci' at the tips of the upper lobes guide the pollinators to the correct part of the flowers and in the correct orientation (Venail 2005, Owen and Bradshaw 2011, Shang *et al* 2011, Tavares *et al* in review).

1.5.4 A hybrid zone between two *Antirrhinum majus* subspecies

Two subspecies of *A. majus* that grow in the Pyrenees have contrasting sets of the colours and guides described above. *A. m. striatum* has bright yellow flowers with dark magenta veins at the centre of the upper lobes, just above the mouth of the flower

(**Figure 1.10**, left). Its sister subspecies, *A. m. pseudomajus*, has magenta flowers with a patch of white on the landing platform ('face') of the flower and yellow foci highlighting the mouth of the flower (**Figure 1.10**, right).

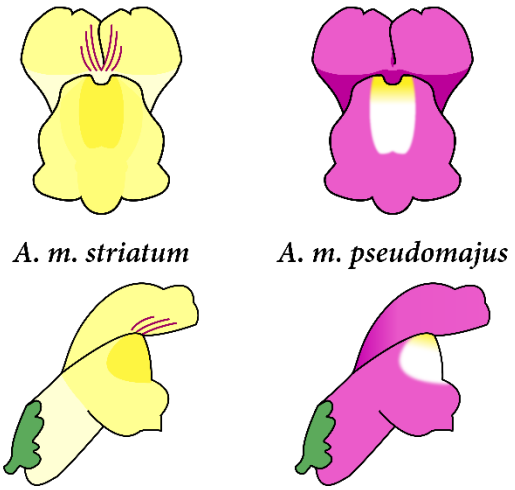


Figure 1.10 Schematic representations of the flowers of *A. m. striatum* and *A. m. pseudomajus*, the former yellow with magenta veins, and the latter magenta with yellow foci.

These contrasting patterns are formed using alternate alleles of several loci of major effect. The *A. m. pseudomajus* allele of *ROSEA* (*ROS*) activates anthocyanin production in the flower lobes (Schwinn *et al* 2006). The *A. m. striatum* allele of *ROS* does not, leaving the lobes unpigmented (Whibley *et al* 2006), apart from in tissue overlying petal veins in the upper lobes, where the same pathway is activated by the *A. m. striatum* allele of *VENOSA* (*VE*) (Venail 2005, Shang *et al* 2011). The effect of *VE* cannot be determined visually in *A. m. pseudomajus* because *VE* is epistatic to *ROS* – the full magenta pigmentation obscures any additional pigmentation in the vein region (Tavares 2014). The *A. m. striatum* allele of *ELUTA* restricts anthocyanin production – and thus the vein pattern – to the centre of the flower (Martin *et al* 1991, Tavares *et al* in review). The *A. m. pseudomajus* allele does not, allowing magenta pigmentation to be unrestricted in the petals. And the *A. m. pseudomajus* allele of *SULFUREA* (*SULF*) restricts aurone production to the foci at the mouth of the flower, which does not happen with the *A. m. striatum* allele (Bradley *et al* 2017). The two subspecies also have different alleles of the aurone biosynthetic gene *FLAVIA* (*FLA*), which encodes CGT (see **Figure 1.7** on page 29). Although both are functional, the *A. m. pseudomajus* allele produces a weaker yellow colour than *A. m. striatum*. This is an alternative method of

restricting yellow in some *Antirrhinum* species and is thought to be redundant in *A. m. pseudomajus* (Boell *et al* unpublished results).

The two subspecies have adjacent population ranges in the south of France and northern Spain/Catalonia, which sometimes overlap. When this happens, hybrid zones are formed. Hybrid zones are found in many plant and animals, with well-studied examples in mice (Turner and Harr 2014), sunflowers (Rieseberg *et al* 1999) and monkeyflowers (Stankowski *et al* 2015). They occur when populations that are genetically distinct meet and produce hybrid offspring. But these areas where hybrids are found are narrow relative to the population ranges of the species, and both populations outside the hybrid zone remain distinct despite gene flow in the contact area (Barton and Hewitt 1989). Studying these hybrid zones can shine new light on the differences between species and the way selection shapes genetic divergence, and as such are described as 'natural laboratories' (Hewitt 1988).

One hybrid zone between *A. m. striatum* and *A. m. pseudomajus* has been studied extensively for more than 15 years (Whibley 2004). This hybrid zone, located in the county of Ripollès in the province of Girona, Catalonia shows a remarkable clinal change in flower colour over a 1-2 km distance (**Figure 1.11 a** and **b**). The plants, being ruderal in habit, grow along two roadsides. Near the village of Fornells de la Muntanya, *A. m. striatum* grows. Travelling east along either one of the roads, *Antirrhinum* flowers with different colours gradually appear: white, orange and pink flowers that are not typically found outside such hybrid zones (**Figure 1.11 c**). Continuing east, another gradual change is seen, with the plants along the roadside near the village of Planoles having magenta flowers – *A. m. pseudomajus*. The phenotypes of the hybrids can be explained using different combinations of the two alleles each at *ROS*, *EL* and *SULF*

Historical evidence suggests that a hybrid zone between *A. m. striatum* and *A. m. pseudomajus* has existed in this region since at least 1928 (Tavares 2014, Tavares *et al* in review). Flower colour plays an important role in maintaining this hybrid zone. In addition to the sharp clines in flower colour across the area, allelic clines and other signatures of strong selection have been described at *ROS* and *EL*, which are linked on one chromosome at the 'ROS-EL locus' (Whibley *et al* 2006, Tavares *et al* in review), and at *SULF* (Bradley *et al* 2017).

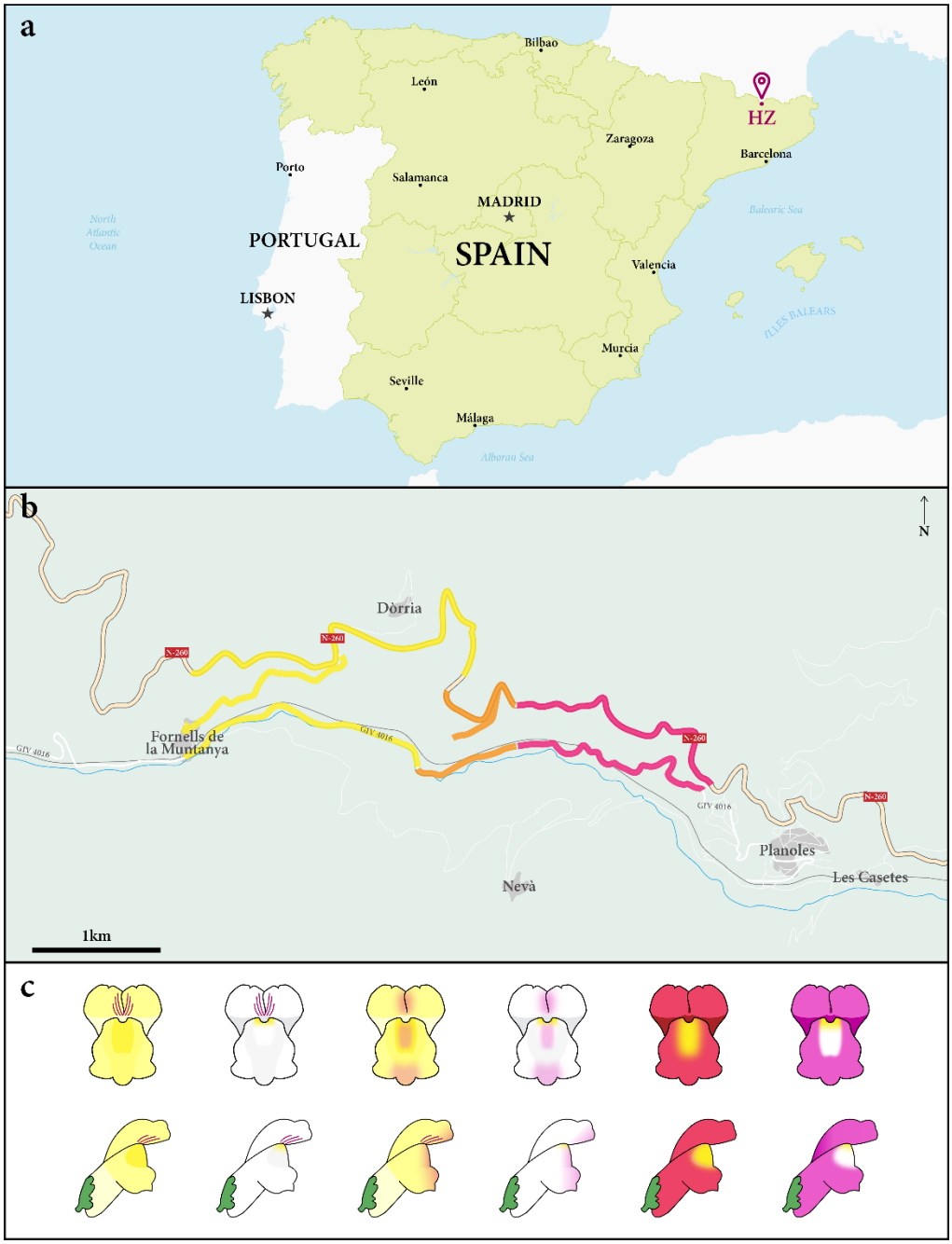


Figure 1.11 Location of the hybrid zone between *A. m. striatum* and *A. m. pseudomajus* on the Iberian Peninsula (**a**) and in the local area (**b**). Also shown are representative phenotypes (**c**). From left to right, along with their *ROS*, *EL* and *SULF* haplotypes, these are: *A. m. striatum* (*ros EL sulf*); white-flowered hybrid (*ros EL SULF*); pale orange-flowered hybrid (*ROS EL sulf*); pink-flowered hybrid (*ROS EL SULF*); bright orange-flowered hybrid (*ROS el sulf*); *A. m. pseudomajus* (*ROS el SULF*).

However, additional phenotypes are seen in the hybrid zone that have not yet been genetically characterised. Although *A. m. pseudomajus*'s magenta pigmentation is known to be regulated by *ROS*, part of the flower is unpigmented. This phenotype is not seen in *A. majus* cultivars where *ROS* also activates anthocyanin pigmentation (Schwinn *et al* 2006). Additional variation has also been generated by crossing plants with hybrid phenotypes from this region to *A. majus*. F2 populations generated from these crosses segregate for traits not seen in either parent. And additional variation in flower colour is also seen in other *Antirrhinum* species, and the genetic loci underlying this variation have yet to be identified.

1.6 Overview and hypotheses

Flower colour in *Antirrhinum* is an ideal system for studying the genetic basis of phenotypic variation in the wild. There a great deal of diversity of colours and patterns between species and populations, and flower colour's tractability as a trait and its tendency to be regulated by loci of major effect makes it easy to identify segregating phenotypes. The interfertility of *Antirrhinum* species means that they can be crossed to *A. majus*, whose genome has been sequenced, so that phenotypes segregate in a known genetic background. And the role of *A. majus* as a model organism gives the advantage that the loci regulating many segregating traits are already known.

In the following chapters, I will present my work testing two alternative hypotheses to explain the genetic determination of flower colour variation in *Antirrhinum*. The first hypothesis is that, given the variability in flower colour seen across the genus, flower colour in *Antirrhinum* is regulated by many different loci – possibly equal to the number of phenotypes that differ between species. The alternative to this hypothesis is that regulation of flower colour is concentrated at a small number of loci, as is seen for wing patterning in *Heliconius*. Different alleles at these loci may be found in species with different flower colours and patterns. I will describe how I used a combination of genetic, genomic and molecular techniques to characterise and map phenotypic differences in anthocyanin pigmentation (chapter 3) and aurone pigmentation (chapter 4) as a means of testing these hypotheses.

2 Materials and methods

2.1 Plant material

2.1.1 *Antirrhinum* accessions

This project used five wild-collected accessions of *Antirrhinum* species from the Iberian peninsula (**Table 2.1**). Collection trips for material used in this work took place in 1999 and 2003, with separate collection trips at a hybrid zone between *A. majus striatum* and *A. m. pseudomajus* in 2009 and 2012. The aim of these trips was to collect seeds from different species with a range of floral phenotypes. Collection codes combine a letter from the alphabet for each year and a three-letter abbreviation for a location. For example, C-NAP was collected in 2003 (year C) near Pont Napoleon in the Pyrenees. Seeds from individual plants were collected and stored separately, with each maternal plant given a unique number. Where possible, 20-30 maternal seed families were collected for a given population and silica-dried leaves were taken for genetic analysis. In addition, for each site, GPS coordinates and habitat descriptions were recorded and photographs were taken. Seeds were stored at 4°C in the John Innes Centre seed store for later use. For seeds collected in the hybrid zone between *A. majus striatum* and *A. m. pseudomajus*, a different numbering system exists because of the collaborative nature, large size and different aims of the hybrid zone project. Plants in the hybrid zone were given a letter corresponding to the year of sampling and a number.

Table 2.1 Wild *Antirrhinum* accessions used in this work. 'HZ' refers to the hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. The first three accessions were collected by members of the Coen lab on collecting trips. For the hybrid zone accessions, location and plant information was collected as part of the hybrid zone project led by Enrico Coen and Nick Barton, and the seeds were collected by Hugo Tavares and Desmond Bradley.

Year collected	Identifier	Plant ID	Species	Latitude	Longitude
1999	Y-GAT	2	*A. charidemi*	36.71°	-2.18°
2003	C-NAP	361	*A. sempervirens*	42.86°	-0.05°
2003	C-QUE	342	*A. molle*	42.11°	1.81°
2009	J (HZ)	1428	*A. m. pseudomajus*	42.32°	2.13°
2012	M (HZ)	0194	*A. m. striatum* × *A. m. pseudomajus* hybrid	42.33°	2.06°

2.1.2 *Antirrhinum* stock cultivars

The John Innes Centre maintains an extensive collection of *Antirrhinum* varieties, mutants and species. Several of these exist as highly-inbred stock lines bred for specific traits – for example, JI7 is a standard *A. majus* stock line with dark magenta flowers, typically used as the wildtype in genetic studies, JI57 is a yellow-flowered mutant and JI659 has radially-symmetrical flowers.

When plants are grown, they are labelled according to growing season (a sequentially changing letter for each season, with one winter and one summer growing season each year), seed family (progeny from a single cross or self) and individual identification number. For example, L124-28 was grown in the summer of 2016 (L) and was the 28th individual in family 124. Some family numbers are reserved for stock lines – families of the line JI7, for example, are always given the number seven.

Another line used in one of my analyses is a $rosea^{dorsea}$ (ros^{dor}) mutant, which has a mutation in the *ROSEA* gene that regulates anthocyanin pigmentation. ros^{dor} was used so that some yellow variation could be seen more clearly and not obscured by anthocyanin pigmentation. This line was originally obtained from the germplasm collection at the Institut für Kulturpflanzenforschung, Gatersleben, Germany (Schwinn *et al* 2006), but has been crossed and backcrossed to JI7 several times to introgress the ros^{dor} allele into the reference genome background (Lucy Copsey, pers

comm).

2.1.3 Plant pedigrees

This project uses crosses between plants grown from wild-collected seed and well-characterised stock lines. Wild accessions for crossing were chosen because of their interesting floral phenotypes discussed in section 2.2.1. The other parent was chosen to have a contrasting phenotype so that progeny would be likely to segregate for the trait of interest. In total, 14 F2 and F3 populations segregating for traits from wild accessions were grown during this project (**Table 2.2**).

Table 2.2 *Antirrhinum* families (populations) used in this thesis. Each family is identified using an identification system (column 1) that combines a letter and number. This system is explained in section 2.1.2 on page 38. Each family shown was grown from the selfed seeds of the parent (column 4). Some families are repeats of families from a previous year (eg L124 is a repeat of J108) – these were seeds from different capsules on the same individual parent. Others are near-repeats (eg N101 and N102 are near-repeats of J108) – these were seeds from sibling parents generated from the same cross. A pedigree of each family is shown in its relevant chapter, and the page number where each pedigree is printed is shown in the final column.

Family ID	Year sown	Family size	Parent	Wild accession	Pedigree page
J108	2015	240	Y135-3 (capsule 1)	J1428 (HZ)	68
L124	2016	500	Y135-3 (capsule 2)	J1428 (HZ)	80
N101	2017	340	Y135-4	J1428 (HZ)	92
N102	2017	320	Y135-5	J1428 (HZ)	92
J104	2015	465	H998-5	C-QUE	100
J152	2015	48	H102R-3 (capsule 1)	M0416 (HZ)	130
J154	2015	48	H102R-15 (capsule 1)	M0416 (HZ)	130
L122	2016	200	H102R-3 (capsule 2)	M0416 (HZ)	140
L123	2016	160	H102R-15 (capsule 2)	M0416 (HZ)	140
N124	2017	168	H102R-2	M0416 (HZ)	144
N136	2017	231	H102R-20	M0416 (HZ)	144
H115	2014	128	E253-8	C-NAP	151
H118	2014	160	E256-2	C-NAP	153
H246	2014	67	D138-7	Y-GAT	155

Crosses were performed manually on emasculated flowers by transferring pollen to the stigma using a paintbrush. Individual flowers were tagged with a paper label and capsules (containing 50-200 seeds) were collected before they dehisced. Initial crosses generated F1 populations. These were then selfed by selecting a flower with a good amount of pollen and brushing it onto the stigma of each flower on the same plant. If not used during the following growing season, seeds were stored at 4°C. In the pedigree diagrams used throughout this thesis, female parents are shown on the left and male parents on the right.

2.1.4 Growth conditions

Seeds were sown by family in trays and watered daily in a lit unheated greenhouse until germination. Seedlings were transplanted into individual 9 cm pots at the one true leaf stage. These were left to grow in the greenhouse and, for summer season plants, transferred outside before flowering. Winter season plants were kept in the glasshouse with 16 hours of supplemental light each day and watered daily.

2.1.5 Harvesting

Flowers were harvested for photographing and phenotyping by gently pulling the pedicle away from the stem. Harvested flowers were individually labelled, kept on ice during collection and stored for a maximum of 12 hours at 4°C to keep them from wilting before phenotyping.

Leaves were harvested in one of two ways. Leaves for individual genetic analysis were collected on dry ice in collection microtubes (Qiagen, Germantown, MD, US) along with a tungsten carbide bead for later grinding, and stored at −80°C until ready for DNA extractions. Leaves for pooled genetic analysis were collected and dried using silica, then stored at room temperature until ready for DNA extractions.

Flower buds were harvested for transcription analysis while they were still developing, when flower colour genes are expressed (Tavares 2014). This corresponded to corolla length between 1 cm and 1.5 cm. Buds were cut using a scalpel and two parts of the flower were harvested: the lateral lobes and the flower face (flower regions shown in **Figure 2.1**). These cut bud parts from several plants with the same phenotype were pooled into an RNase-free 2 ml microcentrifuge tube and flash-frozen in liquid nitrogen. The sample was then stored at −80°C until ready for RNA extractions.

2.1.6 Time considerations

In this thesis, several thousand plants were grown concurrently with leaf and flower material collected from each individual for genotyping and phenotyping work, respectively. Genotyping work needed to be completed before the end of the growing season so that future populations could be obtained by crossing or self-fertilising individuals of interest. This put considerable pressure on the time available for each experiment, limiting phenotyping work to one flower and genotyping work to one leaf sample from each individual plant. This time pressure also meant that the main bulked segregant analysis and individual genotyping experiments were prioritised over gene expression work, which is why RNA sequencing results were not finalised for presentation in this thesis.

2.2 Phenotyping

2.2.1 Visual scoring

Flowers were scored based on visual inspection for a suite of colour traits. These included the intensities of both magenta and yellow pigments, the distribution of colour across the flower and any novel patterns. I chose 10 pattern traits to phenotype in each of the families I looked at (**Figure 2.1**). These were typically easy to score, with pigment (magenta or yellow) accumulating in one or more of the flower regions shown. Individuals could have some combination of these phenotypes, such as magenta veins and yellow foci, although some phenotypes were less distinct, such as overlapping regions obscuring each other. An example of this would be the hinge region, which is masked by the arc region. Initial scoring on small F2 populations was performed on living plants. Once familiarised with the traits, larger populations were grown and flowers were scored from photographs taken in well-lit conditions against a black background with a scale and colour standards. Data were stored in a binary format: presence or absence of each of the 10 traits for each of the two colours.

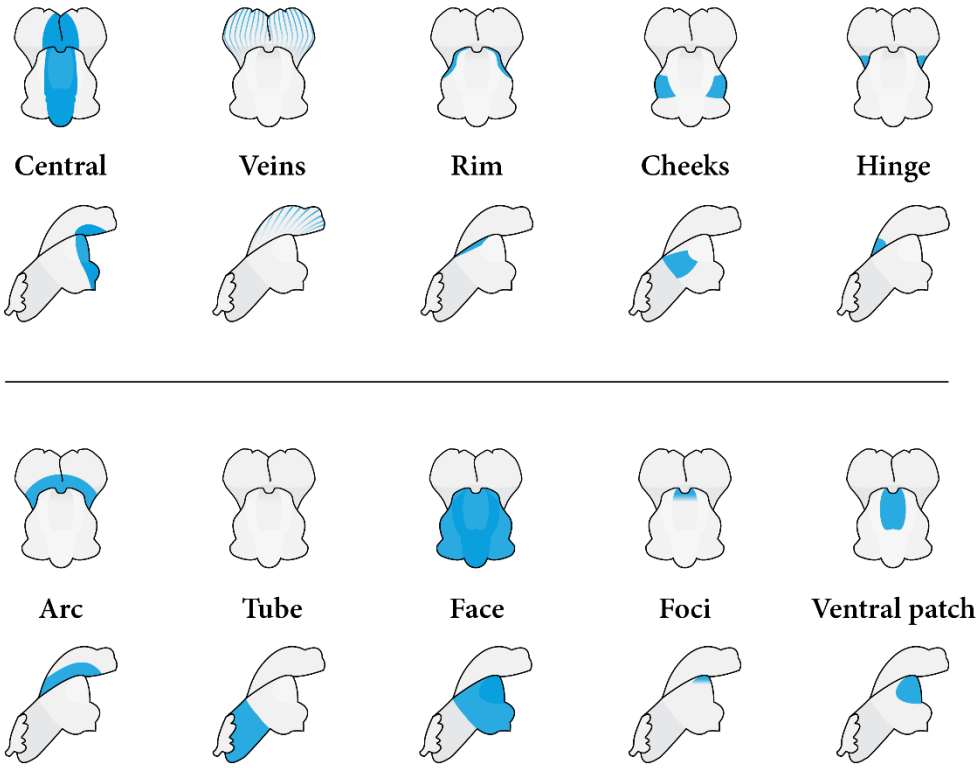


Figure 2.1 Regions of colouration in *Antirrhinum* flowers. Pigments can accumulate in the parts of the flower indicated in blue. Each family was visually inspected for pigmentation differences between individuals in these regions. Segregating families were then phenotyped by individual.

For the traits I looked at in most detail, I developed a more precise scoring system. For the white face trait discussed in chapter 3, this was a numbered system ranging from 0 (no white) to 5 (large white patch). The yellow arc trait discussed in chapter 4 had three discrete forms (little or no yellow, yellow foci, yellow arc, and yellow tube) and therefore a numbering system was not necessary.

2.2.2 Multispectral analysis

Some of the families used in this project had both yellow and magenta pigmentation. Families segregating for the white face trait (chapter 3), for example, also showed variation in the extent of yellow accumulating on the flower face. This made scoring the magenta colour more difficult and tended to introduce uncertainty to the phenotyping. I got around this issue by using a colour-conversion method in Adobe Photoshop. A magenta scoring image was made by systematically converting each

photograph to black and white with magenta pixels darkened. A corresponding yellow scoring image was made by darkening yellow pixels in the same way. Each colour was then scored individually from these colour-converted images. I confirmed that this method was an accurate estimation of the amount of each colour reflected using a multispectral imager (VideometerLab 3, Videometer, Hørsholm, Denmark).

2.3 DNA and RNA isolation

2.3.1 DNA extractions

2.3.1.1 Extraction from silica-dried leaf material

One-to-two young leaves from each sample selected for a pool was placed in a 1.5 ml microcentrifuge tube and disrupted using a micropestle. 400 µl of CTAB extraction buffer (100 mM Tris (pH 8.0); 1.4 M NaCl; 20 mM EDTA (pH 8.0)) was added and the mixture was vortexed. The tube was incubated at 65°C for 30 minutes and then left to cool at room temperature for three minutes. 200 µl of chloroform was added and the tube was vortexed again. The mixture was centrifuged at room temperature at 12,000 rpm for five minutes in a microcentrifuge. 300 µl of the resulting supernatant was transferred to a new tube, to which 200 µl of isopropyl alcohol was added. The tube was inverted several times to mix. This mixture was centrifuged at room temperature at 12,000 rpm for 10 minutes in a microcentrifuge and the supernatant was discarded, leaving a white pellet of DNA. This pellet was washed with 500 µl of 70% [v/v] ethanol, which was then discarded and the pellet was left to air-dry until the pellet was transparent. The DNA was resuspended in 50 µl TE buffer (10 mM Tris-HCl (pH8.0); 1 mM EDTA (pH 8.0)) and quantified using a Qubit broad range DNA assay (Thermo Fisher Scientific, Waltham, MS, US), before being stored at -20°C until it was ready to be sequenced.

2.3.1.2 Extraction from frozen leaf material

Three-to-four young leaves were used for extracting high-quality DNA from individual plants for genotyping work. These leaves were kept frozen at −80°C between harvesting and DNA extraction. DNA extractions were then performed by Richard Goram, who runs a DNA extraction and genotyping facility at the John Innes Centre. Material was ground in a TissueLyser (Qiagen, Germantown, MD, US) and the DNeasy 96 Plant Kit (also by Qiagen) was used for extractions according to the

manufacturer's instructions. Extracted DNA, eluted in sterile water, was kept at −20°C until ready for use. I performed quality control on a subset of around five samples from each plate by quantifying the DNA using a Qubit broad range DNA assay (Thermo Fisher) and measuring absorbance at 260 nm and 280 nm using a NanoDrop 1000 spectrophotometer (also from Thermo Fisher).

2.3.2 RNA extractions

The collected buds were ground to a fine powder in liquid nitrogen using a pre-cooled pestle and mortar. 1 ml of Tri Reagent (Merck Group, Darmstadt, Germany) was then added to each sample and thawed. This mixture was transferred to a 1.5 ml microcentrifuge tube and centrifuged at 4°C at 12,000 rpm for five minutes. The resulting supernatant was transferred to a new microcentrifuge tube and 200 µl of chloroform was added. The mixture was shaken vigorously by hand for 15 seconds before being left to stand at room temperature for three minutes. This mixture was then centrifuged at 4°C at 12,000 rpm for 15 minutes. The upper aqueous layer, containing the dissolved RNA, was transferred to another new microcentrifuge tube and 500µl of isopropyl alcohol was added to precipitate the RNA. This was left to stand at room temperature for 10 minutes and then centrifuged at 4°C at 12,000 rpm for 10 minutes. The supernatant was carefully discarded, leaving a translucent pellet of RNA. The pellet was washed with 1 ml of ethanol, centrifuged at 4°C at 12,000 rpm for 5 minutes, and the ethanol was carefully discarded. The pellet was left to air-dry at room temperature before the RNA was eluted in 20 µl of distilled water treated with diethyl pyrocarbonate (DEPC) to remove RNase enzymes. The isolated total RNA was quantified using a Qubit high sensitivity RNA assay (Thermo Fisher Scientific), absorption was measured as a quality control metric using a NanoDrop 1000 spectrophotometer and the samples were run on an agarose gel to ensure that both 18S and 25S ribosomal RNA bands were present. For this gel electrophoresis, a welled tray of 1.2% agarose gel was prepared with 0.5× Tris/Borate/EDTA (TBE) buffer (45 mM Tris-borate, 45 mM boric acid and 2 mM EDTA) and ethidium bromide was added to a final concentration of 0.4 µg/ml to stain the RNA. The RNA samples were diluted to 1% of their extracted concentrations and a 0.1 volume of loading dye (25% [w/v] Ficoll type 400; 0.25% [w/v] xylene cyanol; 0.25% [w/v] bromophenol blue) was added to each. These samples were loaded into wells and run horizontally alongside a molecular ladder (ssRNA Ladder from New England Biolabs, Ipswich, MS, US) at 10 V/cm until the loading dye's colour could be seen approaching the

end of the gel. The gel was then visualised using a short-wave UV (254 nm) trans-illuminator and photographed.

2.4 High throughput sequencing techniques

2.4.1 DNA sequencing

2.4.1.1 Pooled DNA

DNA was extracted as outlined in section 2.3.1.1. The DNA was pooled in equimolar ratios based on Qubit quantification before being sent for sequencing. DNA sequencing libraries were prepared by The Genome Analysis Centre (TGAC, now the Earlham Institute), Norwich, UK for samples from H115, H118 and H246, and at the Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing 100101, China for all other samples. DNA sequencing was also performed by the same institutes – at TGAC using 100 bp paired-end reads on a HiSeq 2000 sequencer (Illumina, San Diego, CA, US) with a sequencing depth of 20×, and at BIG using 150 bp paired-end reads on a HiSeq 2500 sequencer (Illumina) with a sequencing depth of 50×.

2.4.1.2 DNA from individuals

Although no DNA sequencing of individuals was done for this work, I used previously acquired sequencing data from several *Antirrhinum* species. These resequenced genomes were based on short read Illumina data, sequenced by TGAC and mapped to the *Antirrhinum* genome by Annabel Whibley.

2.4.2 The *Antirrhinum* genome

The reference *Antirrhinum* genome was sequenced and assembled by Yongbiao Xue and is based on short read Illumina data and longer Pac Bio (Pacific Biosciences, Menlo Park, CA, US) reads.

2.4.3 Bioinformatic analysis

2.4.3.1 Quality testing

Sequencing reads in FASTQ format were first quality-tested using `fastq-mcf`, part of

the ea-utils suite of bioinformatics software (Aronesty 2011). The following commands were used to process the FASTQ reads:

```
1  fastq-mcf -l 50 -k 0 -x 20 -q 20
   --qual-mean 20
   --qual-gt 90,20
   --max-ns 0
   --min-len 50 n/a
   -o <filtered_fastq_read1>
   -o <filtered_fastq_read2>
   <read1_fastq> <read2_fastq>
```

2.4.3.2 Mapping to the reference genome

All reads were mapped to the *Antirrhinum* reference genome (version 2) using the Burrows-Wheeler Aligner (BWA) (Li 2013) using these commands:

```
2  bwa mem <reference_genome>
   -R <read_group_info>
   <read1_fastq> <read2_fastq>
   > <output_SAM_file>
```

2.4.3.3 Processing

Following mapping, the files were processed using a combination of Picard Tools (Broad Institute 2018b), SAMtools (Li *et al* 2009) and the Genome Analysis Toolkit (GATK, version 3.6) (McKenna *et al* 2010, Broad Institute 2018a). SAMtools was used to convert the BWA output sequence alignment/map (SAM) files to the binary alignment/map (BAM) format accepted by downstream processing software:

```
3  samtools view -h -b <SAM_file> -o <output_BAM>
```

I then sorted the contents of the BAM file, again using SAMtools:

```
4  samtools sort <input_BAM> -o <output_BAM>
```

Duplicate reads, which can arise during PCR amplification of sequencing libraries or through incorrect detection of single reads as multiple ones, were removed using Picard Tools' `MarkDuplicates` tool:

```
5  java -Xmx32g -jar MarkDuplicates.jar
   INPUT = <sorted_BAM>
   OUTPUT = <BAM_without_duplicates>
   REMOVE_DUPLICATES=true
   ASSUME_SORTED=true
   METRICS_FILE = <output_metrics_file>
   MAX_FILE_HANDLES_FOR_READ_ENDS_MAP = 1000
   CREATE_INDEX = true
```

To minimise the number of mismatching bases in the aligned sequencing file, GATK tools were used to locally realign reads in intervals. The intervals for this realignment

were created using `RealignerTargetCreator` and the reads were realigned using `IndelRealigner`:

```
6  java -Xmx16g -jar GenomeAnalysisTK.jar
   -T RealignerTargetCreator
   -R <reference_genome>
   -I <BAM_without_duplicates>
   -o <realignment_intervals>

7  java -Xmx16g -jar GenomeAnalysisTK.jar
   -T IndelRealigner
   -R $ref.fasta
   --targetIntervals <realignment_intervals>
   -I <BAM_without_duplicates>
   -o <realigned_BAM>
```

2.4.4 Bulked segregant analysis

2.4.4.1 SNP/indel calling

Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) were called using `mpileup` from SAMtools:

```
8  samtools mpileup
   -C 50 -q 40 -Q 30
   -f <reference_genome>
   -o <mpileup_output>
   <realigned_BAM>
```

The second line contains commands for quality filtering. Here, `-C` is a coefficient for downgrading the mapping quality score given to reads that contain excessive mismatches. A value of 50 is recommended for reads mapped using BWA. The `-q` command sets a minimum coefficient-adjusted mapping quality to use; reads not meeting this threshold are not used. The `-Q` command sets a minimum base quality threshold. The output was given in a `.pileup` format.

2.4.4.2 Allele frequency calculation

Frequencies of the reference (JI7) and non-reference alleles in the pools at each SNP were estimated using SNAPE-pooled (Raineri *et al* 2012). SNAPE-pooled uses a Bayesian estimation of SNP posterior frequency distribution in pooled samples. The software is intended for pools of wild individuals which would show more diversity within the pools than my F2 individuals. For this reason, there are a number of parameters to specify that differ from the default values suggested by the authors.

```
9  snape-pooled
   -nchr <number_of_alleles>
   -theta 0.05
   -D 0.05
   -priortype flat
   -fold folded
   < <mpileup_output>
   > <snape output>
```

The `-nchr` command specifies the number of alleles expected at each SNP – in *Antirrhinum*, 2× the number of individuals in the pool. The `-theta` value specifies the nucleotide diversity (θ) in the pool. A value is required but will be ignored by the program because the `-priortype` is set as `flat`, meaning that we do not make assumptions about the allele frequencies in the pools. `-D` is the prior genetic difference between the reference genome and the individuals in the pool. This is very low in these analyses as the reference genome used is based on JI7, which was one of the parents of each cross used to generate the F2s. The `folded` value for the `-fold` parameter is used because we do not know the identity of the non-reference allele.

2.4.4.3 Calculating G and G' statistics

A method for statistical analysis of genomic BSA was developed by Magwene *et al* (2011). This method starts by combining the allele counts of both pools to give four values for each SNP: count of the reference allele in pool 1 (n_1), count of the non-reference allele in pool 1 (n_2), count of the reference allele in pool 2 (n_3), and count of the non-reference allele in pool 2 (n_4). A modified G statistic, based on the standard G statistic (equation 1) is calculated for each SNP based on these allele counts. This uses an expected value of n_i ($\hat{n}_i$) based on the null hypothesis that there is no causal locus near a SNP (equation 2). For SNP where this null hypothesis is correct, and assuming no distortion in segregation or average sequencing coverage, $\hat{n}_1 = \hat{n}_2 = \hat{n}_3 = \hat{n}_4$.

$$G = 2\sum_{i=1}^{4} n_i \ln\left(\frac{n_i}{\hat{n}_i}\right) \qquad (1)$$

$$\hat{n}_i = \frac{(n_1 + n_2)(n_1 + n_3)}{n_1 + n_2 + n_3 + n_4} \qquad (2)$$

However, because BSA sampling is in two phases – ie first the individuals are divided

into two phenotypic pools, and then two alleles are sampled in each pool – G is not expected to follow the usual χ_1^2 distribution. Therefore, Magwene *et al* (2011) developed an alternative simulation to calculate a modified G, based on the expected distribution in BSA studies. A smoothed version of the G statistic (G') is also calculated for each SNP, which involves averaging G across a window (W) using a weighted kernel regression (k) (equation 3). For my analyses, I used window sizes of 50 kb. As a kernel function, I used the tri-cube kernel recommended by Magwene *et al* (2011), which takes into account the standardised distance within the window (0 at the focal SNP, 1 at the edge of the window) (equation 4).

$$G' = \sum_{j \text{ in } W} k_j G_j \qquad (3)$$

$$k_j = \frac{(1 - D_j^3)^3}{S_W} \qquad (4)$$

2.4.4.4 Running and plotting analyses

A pipeline for carrying out these analyses in R (R Development Core Team 2008) was developed by Mansfeld and Grumet (2017). I made small modifications to these scripts to allow smaller window sizes. Their scripts also calculate a Δ SNP-index, an estimation of the allele frequency difference between the pools, which was developed by Takagi *et al* (2013). These were also smoothed using the tri-cube kernel function.

I plotted the number of SNPs in each window, genome-wide G' values and genome-wide smoothed Δ SNP-index values against physical positions along each chromosome according to version 2 of the *Antirrhinum* genome (Xue *et al*, in preparation) using ggplot2 (Wickam 2009) and compiled them for the figures presented in this thesis using Adobe Illustrator (Adobe Systems, San Jose, CA, US).

2.4.5 RNA sequencing

Extracted RNA was sent to BIG for preparation of Illumina libraries and RNA sequencing. This work is ongoing, and the results of these analyses were not completed in time for presentation in this thesis.

2.5 SNP genotyping

2.5.1 KASP genotyping

Kompetitive Allele-Specific PCR (KASP) is a technology developed by LGC Genomics, Teddington, UK to detect the presence of one or both of a pair of alleles at a specific SNP. Two oligonucleotide primers are designed in one direction, one ending in the base pair of each allele at the focal SNP (**Figure 2.2**). A common primer in the opposite direction is used to amplify a 20-100 bp PCR product. For this explanation I will consider the allele-specific primers to be in the forward (5′ to 3′ direction) and the common primer to be in the reverse direction. Each of the allele-specific primers has a fluorophore-labelled oligonucleotide tail at its 5′ end, with a VIC® cassette-targeting sequence for one allele and a FAM™ cassette-targeting sequence for the other allele (LGC Ltd 2013).

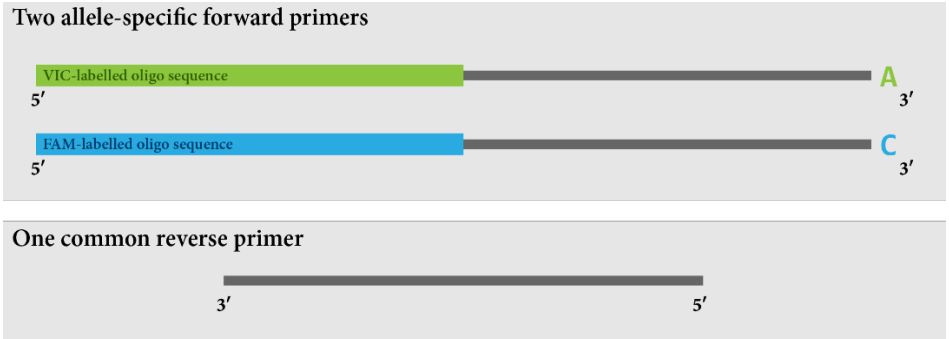


Figure 2.2 Examples of the three primers used in KASP genotyping. The two allele-specific primers (in the forward direction in this case) start with a fluorophore-labelled primer sequence (VIC for one allele, FAM for the other) and end with the base pair of their respective alleles at the focal SNP. The common reverse primer matches a sequence in the sample 20-100 bp downstream of the focal SNP.

The KASP reaction mixture contains the fluorescence cassettes targeted by the oligonucleotide tails, a Taq polymerase enzyme, a buffer solution and DNA from the sample being tested, at a total volume of 10 µl in each well of a 96-well plate, although lower volume reactions in plates with up to 1536 wells are possible (LGC Ltd 2013). The reaction is then run on the programme specified by the company: 94°C for 15 minutes to initiate the hot-start Taq polymerase enzyme, 10 cycles of 94°C for 20 seconds and 61°C for 60 seconds (decreasing by 0.6°C in each subsequent cycle), and then 26 cycles of 94°C for 20 seconds and 55°C for 60 seconds. I ran these plates in a standard 96-well PCR thermocycler and read the plates on a CFX96 Touch™ real-

time PCR detection thermocycler (Bio-Rad laboratories, Hercules, CA, US). Fluorescence in the blue (for FAM) and green (for VIC) channels was measured in relative fluorescence units (RFU) and genotypes of reactions were called from the way they clustered.

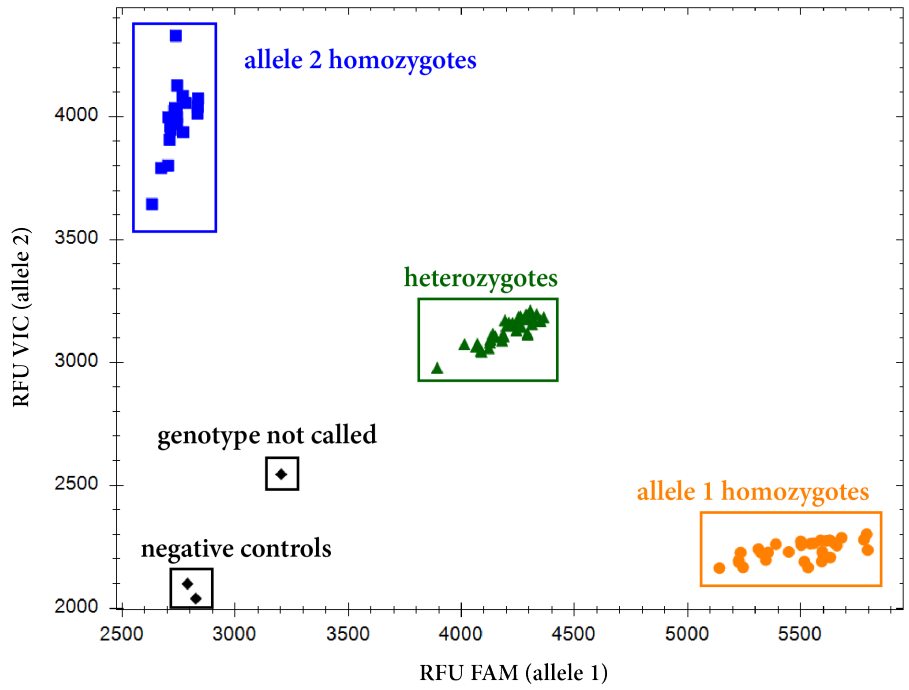


Figure 2.3 Example output from the CFX Manager 3.1 (Bio-Rad Laboratories) software used to call the genotypes of KASP reactions. Reactions clustering together that had high FAM fluorescence and low VIC fluorescence were called as allele 1 homozygotes and vice versa for allele 2. Reactions clustering together that had high fluorescence for both fluorophore channels were called as heterozygotes. Individuals not clustering with others were not called. These may represent wells where the DNA was not correctly extracted or amplified. At least two negative controls, one only containing the reaction mixture and water and one containing the reaction mixture plus a negative DNA extraction control (a well left empty during collection, but that was treated as a sample during extraction).

2.5.2 Marker determination

Oligonucleotide primers for KASP genotyping were developed in one of two ways. Firstly, I designed several primers myself using the Primer3 (Koressaar and Remm 2007, Untergasser *et al* 2012) plugin in Geneious (Kearse *et al* 2012). These sequences were based on the sequencing data from bulked segregant analyses to identify SNPs that were surrounded by non-polymorphic regions long enough to accommodate a 22-base primer sequence. The allele-specific primers ended in the base seen at the focal

SNP in the 5′ to 3′ strand in each of the two pools. The following tails were attached to the 5′ end of these oligonucleotide sequences: GAAGGTCGGAGTCAACGGAT for primers targeting the VIC cassette; and GAAGGTGACCAAGTTCATGC for primers targeting the FAM cassette. These primers were ordered from Sigma-Aldrich Life Science (Merck Group, Darmstadt, Germany). Secondly, some primers were designed by LGC Ltd using their proprietary software. For these primers, the sequence surrounding the focal SNP was sent to LGC and the primers were received ready to use. The LGC-designed primers were acquired through collaboration with David Field (University of Vienna, Austria) and Nick Barton (Institute of Science and Technology Austria, Klosterneuburg, Austria).

3 Transcription factors regulate magenta colour variation in *Antirrhinum majus*

3.1 Introduction

Crucial to developmental processes in plants and animals is the regulation of gene expression. A common mechanism for regulating where and when genes are expressed in plants is through transcription factors. These are proteins that interact with *cis*-regulatory sequences in the promoter region of a gene to regulate transcription of that gene (Meshi and Iwabuchi 1995). There are several classes of transcription factors in plants, classified according to their conserved domains – regions that are highly conserved within the protein families and that bind a target DNA sequence within the *cis*-regulatory regions of target genes (Liu *et al* 1999).

As discussed in chapter 1, the genetic basis of phenotypic variation can be due to genetic differences that are localised to one part of the genome (centralised genetic control) or mutations in different parts of the genome that each affect a given trait in a different way (dispersed genetic control). In the case of trait regulation by transcription factors, centralised control could mean a single transcription factor regulating a phenotype in different populations, possibly with different effects. An example of this is wing patterning mediated by a transcription factor in *Heliconius* butterflies, where several species have red bands on their wings as warnings to predators, with the same

genes used in different species to produce different banding patterns (Baxter *et al* 2008, Reed *et al* 2011). Dispersed control could mean several transcription factors encoded by genes dispersed throughout the genome each affecting the same trait. Height in humans is a polygenic trait controlled by hundreds of loci, many of which encode transcription factors (Lettre *et al* 2008, Weedon *et al* 2008, Simeone and Alberti 2014).

3.1.1 Flower colour patterns in *Antirrhinum*

Antirrhinum flowers are insect-pollinated and, like other insect-pollinated species, they have evolved to attract pollinators to visit them using specialised cues. One of the most striking of these cues in *Antirrhinum* is the colours and patterns produced on the flowers. In this chapter, I will focus on the magenta colour seen, with varying degrees of intensity and patterning, in all *Antirrhinum* species. In particular, I will describe my efforts to characterise the 'white face' phenotype found in *A. m. pseudomajus* and other species with predominantly magenta flowers.

3.1.2 Regulation of magenta colour in *Antirrhinum* petals

Antirrhinum flowers' magenta colour is caused by anthocyanin pigments, a class of flavonoid secondary metabolites found in nearly all land plants (Campanella *et al* 2014). Anthocyanin pigments accumulate in the petals' upper epidermal cells and, under acidic conditions, absorb light at a range of wavelengths according to the number of hydroxyl groups on one of the rings of the molecule – one in pelargonidin, two in cyanidin and three in delphinidin (Glover and Martin 2012). In *A. majus*, the 3-rutinoside of cyanidin that gives the flowers their characteristic hue, was isolated by Scott-Moncrieff (1930), although some cultivars also produce cyanidin 3-glucoside in addition to the 3-rutinoside, resulting in darker petals (Gilbert 1971). Cyanidin has an absorption maximum between 465 nm and 550 nm (green light) and a secondary absorbance peak between 270 nm and 280 nm (ultraviolet light); pelargonidin's main peak is wider than that of cyanidin and its absorption maximum is at a shorter wavelength, so it absorbs some blue light (400-450 nm) in addition to green (Saito and Harborne 1992) (**Figure 3.1**). These absorbance profiles mean that cyanidin is seen by humans as magenta (a combination of red and blue-violet), while pelargonidin appears red (Gausman 1983, Glover and Martin 2012). Many pollinators, including bees and hummingbirds, use colour to forage, although their visual spectra – the range of wavelengths their photoreceptors can detect – can vary (Osorio and Vorobyev 2008).

Anthocyanins are synthesised in the cytoplasm and the endoplasmic reticulum through the flavonoid biosynthesis pathway. The pigment molecules are then deposited in the vacuole (Zhao and Tao 2015).

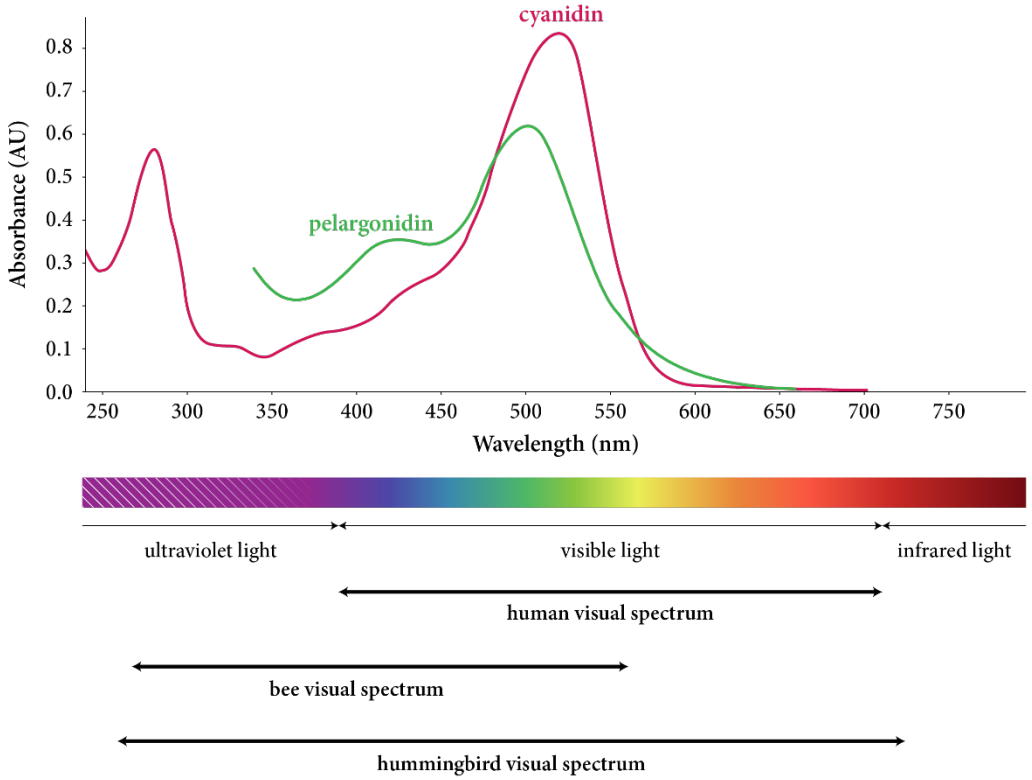


Figure 3.1 Absorbance spectrum of cyanidin 3-glucoside as measured by Skaar *et al* (2014) and pelargonidin 3-glucoside as measured by Lopes-da-Silva *et al* (2007). Absorbance is shown in absorption units (AU). Shown below the graphs are the part of the electromagnetic spectrum for the wavelengths along the *x*-axis and the visual spectra of humans, bees and hummingbirds.

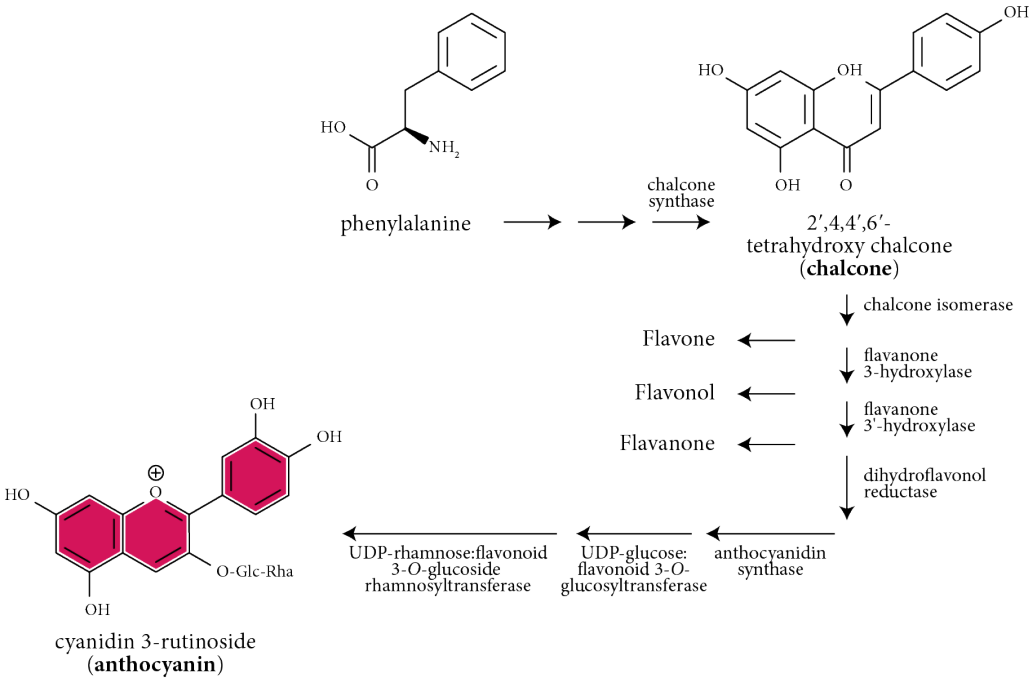


Figure 3.2 Simplified biosynthetic pathway of cyanidin 3-rutinoside, the magenta-coloured anthocyanin found in *Antirrhinum majus* and related species. The main steps are shown in the conversion of a molecule of the amino acid phenylalanine to the final form of cyanidin found in the vacuoles of magenta-pigmented *A. majus* epidermal cells. The intermediate stages between chalcone and cyanidin also give rise to other important flavonoids, such as flavones and flavonols. Adapted from Falcone Ferreyra *et al* (2012).

At least eight enzymes are required to make anthocyanins in plants. Most of the genes encoding them have been identified by studying mutants. These biosynthetic enzymes work consecutively to convert a molecule of the amino acid phenylalanine and three molecules of malonyl CoA to make an anthocyanin molecule (**Figure 3.2**). The intensity of the pigment in different cells depends on the expression of the genes that encode the enzymes, which in turn is controlled by the expression of regulatory genes encoding transcription factors (Martin *et al* 1991, Schwinn *et al* 2006, Streisfeld and Rausher 2011).

The anthocyanin pathway in *Antirrhinum* is regulated by MYB-like proteins – typically comprising two repeat sequences containing R2 and R3 MYB motifs, each of which encodes α-helices that bind to DNA (Martin and Paz-Ares 1997, Stommel *et al* 2009). The R2R3 MYB-like transcription factor-encoding regulatory genes *ROSEA* (*ROS*), *ELUTA* (*EL*) and *VENOSA* (*VE*) are all known to regulate anthocyanin production (Schwinn *et al* 2006, Shang *et al* 2011, Tavares *et al* in review). The transcription factors

they encode interact with late stages of the anthocyanin synthesis pathway, controlling the presence or absence of the pigment in different petal regions (**Figure 3.3**). In magenta-flowered *Antirrhinum* species, the dominant allele of *ROS* switches on anthocyanin production in epidermal cells throughout much of the flower face, resulting in brightly coloured flowers. In *ROS*/- plants where the functional semidominant *EL* allele is homozygous, anthocyanin production is limited to the centre of the flower, a phenotype not seen in the wild outside naturally occurring hybrid zones. Typically, *Antirrhinum* species with a *ROS* genotype have a non-functional *el* copy, which means that magenta colouration is not restricted. The third gene *VE* switches on anthocyanin production in epidermal cells overlying veins in the dorsal lobes of the *Antirrhinum* flower, thus producing a radiating magenta venation pattern throughout this region. In plants with the *VE*/-, *EL*/*EL* genotype, these vein patterns are localised to the centre of the flower.

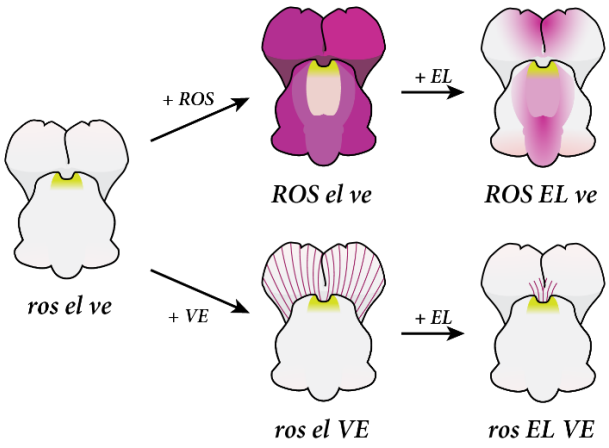


Figure 3.3 The effects of the MYB-like transcription factor-encoding genes *ROSEA* (*ROS*), *ELUTA* (*EL*) and *VENOSA* (*VE*) on magenta flower colour in *A. m. pseudomajus*. All genes are considered homozygous in this diagram. Starting with a *ros el ve* triple mutant, the addition of *ROS* gives rise to magenta pigmentation throughout much of the petals. With the addition of *EL*, this pigmentation is restricted to the centre of the flower. From the same starting triple mutant phenotype, adding *VE* gives a flower phenotype with veins across the dorsal petals. Adding *EL* restricts these veins to the centre of the flower. The effect of *VE* cannot clearly be seen in a *ROS* background as the bright petal-wide pigmentation obscures colour in the veins.

The *ROSEA* locus can be further divided into three genes, *ROS1*, *ROS2* and *ROS3*. These have similar DNA sequences and are thought to have arisen through duplication of an ancestral *ROS* locus. ROS1 has been shown to activate flavanone 3-

hydroxylase (F3H), flavanone 3′-hydroxylase (F3′H), dihydroflavonol reductase (DFR) and UDP-glucose: flavonoid 3-*O*-glucosyltransferase (UF3GT), while ROS2 activates only F3′H (Schwinn *et al* 2006, Glover 2014). The *rosea*^{dorsea} (*ros*^{dor}) mutant line of *A. majus* has vastly reduced anthocyanin pigmentation. Gene expression studies have shown that this line lacks *ROS2* expression and has changes in the promoter region of *ROS1*. Another mutant line, *rosea*^{colorata} (*ros*^{col}), also has reduced anthocyanin production; *ROS2* is expressed in this line, but *ROS1* is not (Schwinn *et al* 2006). *ROS3* has only recently been described through sequence analysis around the *ROS* locus. No *ros3* mutants have been characterised and it may be a pseudogene (Tavares 2014, Tavares *et al* in review).

EL, located just 100 kb from the *ROS* locus, likely arose through a duplication event at *ROS*. The *EL* locus is predicted to be 609 bp long and encodes a transcription factor with two MYB domains. When the *ros EL* and *ROS el* haplotypes from other *Antirrhinum* species are crossed to a common *A. majus* background, expression of the *EL* gene is significantly different between *ros EL* and *ROS el* plants (**Figure 3.4**) (Tavares *et al* in review).

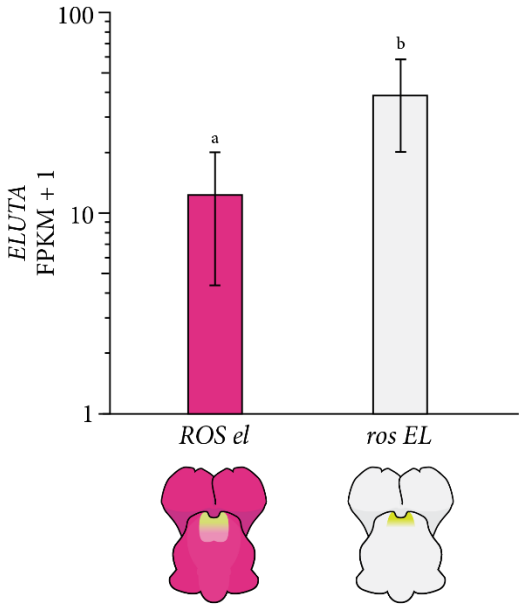


Figure 3.4 Expression estimates *EL* in the buds of whole flowers with different genotypes at *ROS* and *EL*. The bars show the mean and 95% confidence interval of expression, measured as fragments per kilobase of transcript per million mapped reads (FKPM). The flower illustrations depict the phenotype seen for each haplotype. Expression estimates are significantly different between plants with each haplotype ($q < 0.01$). Adapted from Tavares *et al* (in review).

3.1.3 White face phenotype in the *Antirrhinum* genus

A common feature of most *Antirrhinum* species with magenta flowers is a lack of anthocyanin pigmentation in the face region around the flower foci – the two points at the top of the lower lobes (**Figure 3.5**). In some species, such as *A. majus pseudomajus*, this lack of magenta extends through much of the face of the flower, whereas in others, such as *A. linkianum*, it is confined to the few millimetres immediately surrounding the foci or not present at all (**Figure 3.6**). These, however, are generalisations and much variation in this 'white face' phenotype also exists between populations of the same species. Several cultivars of *A. majus* also have magenta flowers, and many of these lack the white face phenotype. *A. majus* var. JI7, a research line commonly used for genetic studies of flower colour, is one such cultivar that has full magenta colouration.

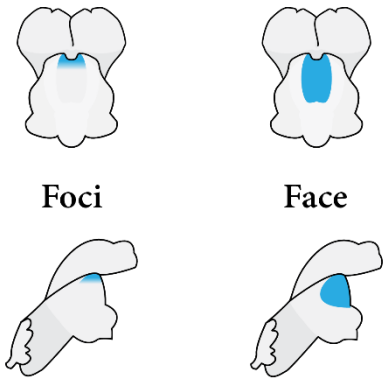


Figure 3.5 An illustration of the locations on *Antirrhinum* corollas of the regions referred to as the 'foci' and the 'face'.

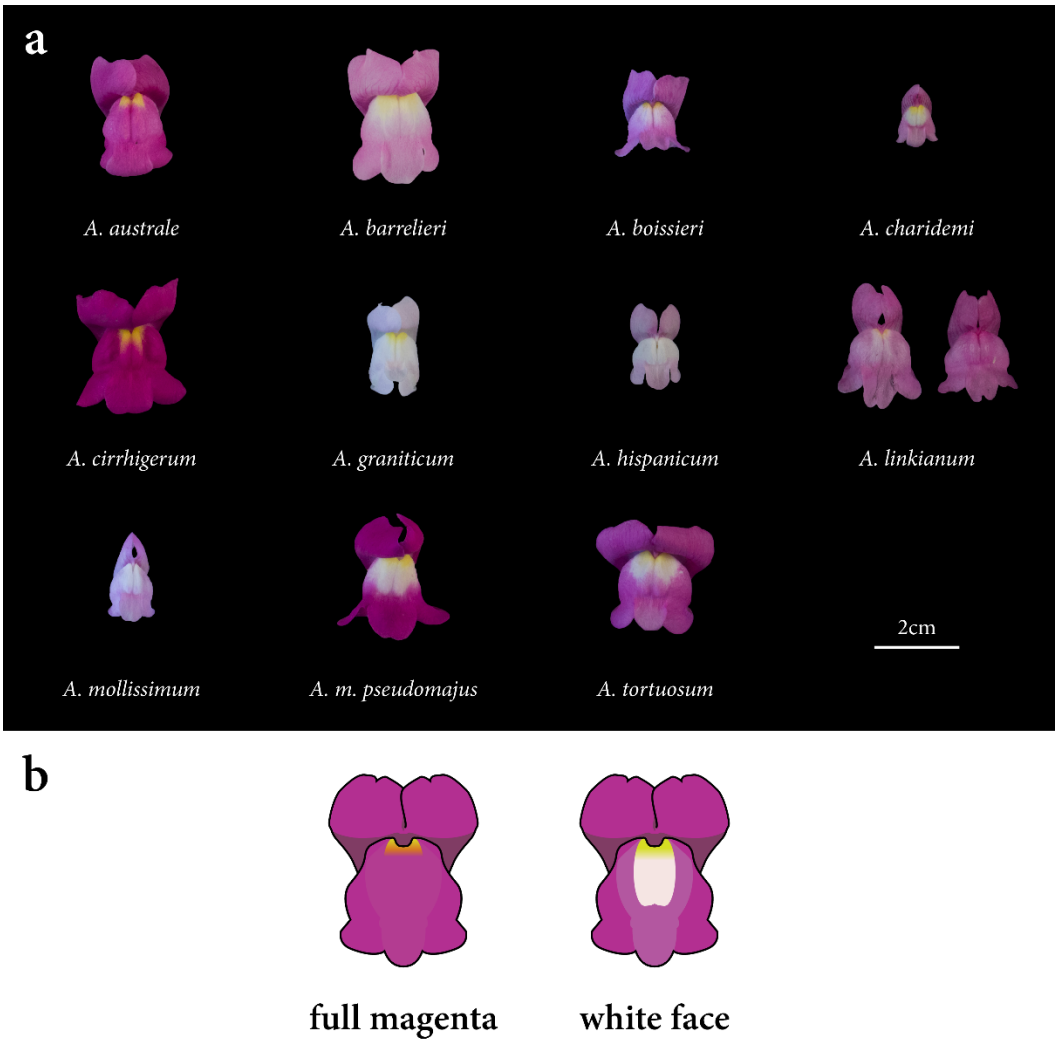


Figure 3.6 Photographs of the flowers of 11 *Antirrhinum* species with pink or magenta flowers (**a**) and schematic representations of the full magenta and white face phenotypes (**b**). In some species, such as *A. m. pseudomajus* and *A. barrelieri*, a large part of the flower face is left white owing to a lack of anthocyanin pigmentation. In others, such as *A. cirrhigerum* and *A. australe*, this region without magenta is generally smaller and only includes the foci, where yellow aurones accumulate. In *A. linkianum*, some accessions have no white patch on their flowers.

Another species of interest in regard to the white face phenotype is *A. molle*. This species, which grows in southwestern France, northern Catalonia and Andorra (**Figure 3.7**), has white flowers, with no magenta outside the dorsal vein pattern seen in many white-flowered *Antirrhinum* species. However, when crossed to the lab cultivar JI7, a white face pattern has been observed in F2 populations. This may mean that the white face allele of *A. m. pseudomajus*, whose habitat range neighbours that of

A. molle, is also seen in other species but does not always produce a phenotype because of epistasis. Alternatively, the white face phenotype seen in these F2 populations may have an entirely different genetic basis.

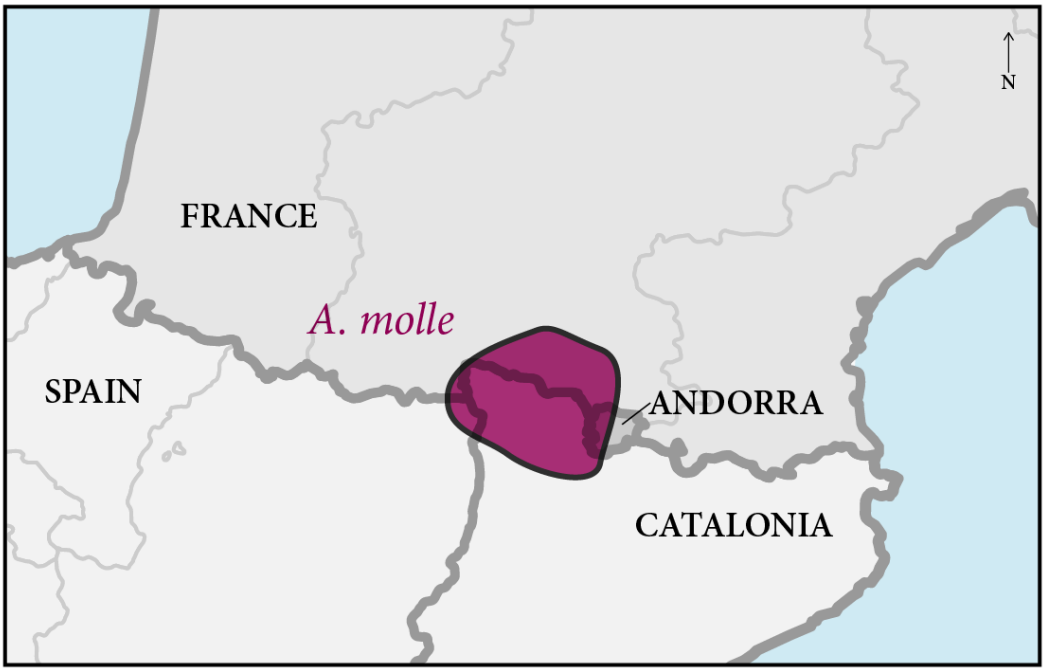


Figure 3.7 Approximate population range of *A. molle* near the borders between France, Catalonia and Andorra. Drawn using information from Whibley (2004) and Wilson and Hudson (2011).

Figure 3.8 Images of *Antirrhinum molle* and its flowers. Front view of a flower in the wild (**a**), side view of a flower from the same plant in the wild (**b**) and growth habit of a plant in the wild (**c**) – all taken at the QUE location near the Santuari de Queralt in Berga, Catalonia, in 2017, and thus given the identifier W-QUE. Front (**d**) and side (**e**) view flower photographs from a plant germinated from seed collected at the same location in 2003 (identified as C-QUE).

The adaptive advantage, if any, of this white face phenotype has not been established, nor has its effectiveness at attracting pollinators been characterised. But one explanation may be that clearing the foci and surrounding regions of anthocyanins ensures that different pigments do not accumulate in the same parts of the flower. The foci of predominantly magenta-flowered *Antirrhinum* species accumulate a yellow pigment called aurone, and these are thought to serve as nectar guides to facilitate pollination. A similar yellow pattern is found in *Mimulus lewisii* (Phrymaceae) and mutants lacking these yellow guides receive around 20% fewer bumblebee visitors than wildtype plants. Also, 55% of visitors entered the mutant flowers in the wrong orientation, compared with 10% of visits in the wrong orientation on wildtype flowers (Owen and Bradshaw 2011) (**Figure 3.9**). Allowing the magenta pigment to encroach on this yellow region may dilute the signal these foci conveys to pollinators, thus reducing the plant's reproductive success.

Figure 3.9 Bumblebees foraging on wildtype and mutant *Mimulus lewisii* flowers. Bumblebees enter wildtype flowers facing upwards, guided by yellow foci on the petals (a). These guides are missing in the *guideless* mutant, resulting in 55% of pollinators entering in the wrong orientation (b). The same entry orientation is seen when wildtype flowers are turned upside down (c). Adapted from Owen and Bradshaw (2011).

3.1.4 Similar phenotypes in other plant species

Antirrhinum is not the only system where a flower colour pattern is conferred by localised lack of pigment. In the bumblebee-pollinated monkeyflower *Mimulus lewisii*, the petals are coloured pink by anthocyanins, with yellow foci coloured by carotenoids (**Figure 3.10 a**). However, the region at the 'throat' of the flower, surrounding the yellow foci, is left white in a similar manner to the white face of *A. m. pseudomajus*. This phenotype in *M. lewisii* has been mapped to a single Mendelian locus called *LIGHT AREAS1* (*LAR1*). This locus encodes an R2R3-MYB transcription factor,

which targets the flavonol biosynthetic gene *FLAVONOL SYNTHASE* (*FLS*), and plants carrying the dominant allele have higher *FLS* expression than those homozygous for the recessive *lar1* allele (**Figure 3.10 b**) from the related species *M. cardinalis* (**Figure 3.10 c**). The flavonol synthase enzyme encoded by *FLS* produces a colourless flavonol that competes with, and eliminates, anthocyanin biosynthesis in the light region of the flower (Yuan *et al* 2016). Similar transcription factors to that encoded by *LAR1* are important for flower colour in *A. majus* – *ROS*, *EL* and *VE* all encode R2R3 MYBs that regulate anthocyanin biosynthesis (Schwinn *et al* 2006). No *LAR1* homologue has been described in *Antirrhinum*, but one possible explanation for the white face phenotype is that flavonol and cyanidin biosynthesis are in competition in the white face region of *Antirrhinum* flowers. Alternatively, *Antirrhinum* may have independently evolved a similar phenotype using a different mechanism. This can be tested by mapping the gene behind this phenotype and determining its function in *Antirrhinum* flowers.

Figure 3.10 The flowers of *Mimulus lewisii*, *M. cardinalis* and a *lar1/lar1 M. lewisii* near-isogenic line (NIL). Wildtype *M. lewisii* has a white patch on its flowers' throat region. Its close relative *M. cardinalis* lacks this white patch; its flowers are also dark red, unlike *M. lewisii*'s pink flowers. This white patch is caused by the accumulation of flavonol in the throat region, which reduces the amount of pink anthocyanin accumulated. This flavonol accumulation is regulated by the LAR1 MYB-like transcription factor in *M. lewisii*. In the *lar1* NIL, anthocyanins accumulate throughout the throat region. Photographs from Yuan *et al* (2016)

3.1.5 Pooled whole genome sequencing

Determining the link between phenotype and genotype – ie determining which genes affect which traits – is a core aim in genetics. A widely used and simple method for mapping genetic variation in this way is bulked segregant analysis (BSA), a technique first developed by Michelmore *et al* (1991) for the detection of phenotype-associated

genetic marker sites. This method involves generating a population that segregates for a trait of interest and pooling individuals into two bulks according to that trait – one bulk containing individuals with one extreme phenotype and another containing those with the contrasting extreme phenotype. This can be done either for traits that appear to segregate in a Mendelian fashion – ie one gene linked to the phenotype, producing a 3:1 phenotypic ratio – or for more complex traits regulated by several quantitative trait loci (QTL). All individuals in the same bulk will have the same phenotype for the trait being studied. When allele frequencies in pooled DNA samples from each bulk are estimated and the difference between them calculated, one would expect a marker linked to the trait of interest to show high frequency differences between the bulks, while markers at other sites in the genome would have similar frequencies in both bulks. Thus, a peak in allele frequency difference will reveal a locus linked to the trait of interest (**Figure 3.11**). The higher the number of recombination events and more accurate the estimation of allele frequency, the more effective the analysis and the narrower these peaks will be (Magwene *et al* 2011).

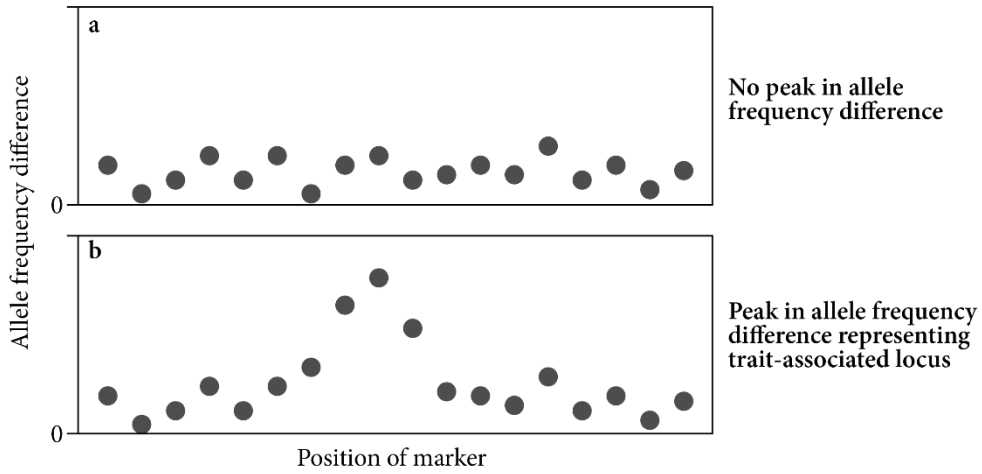


Figure 3.11 Representation of the results from two hypothetical bulked segregant analyses, each using 18 markers. In the first analysis, there is no locus associated with the trait that differs between the pools, and all markers show little difference in allele frequency estimates (**a**). In the second, there is a locus associated with the trait, and markers linked to that locus show high differences in allele frequency between the pools (**b**).

Advances in the technology used for DNA sequencing during the last decade have made whole genome sequencing (WGS) a powerful and effective tool both in evolutionary biology and genetics. Genome resequencing (ie sequencing the genome of a species for which a reference genome already exists) has particularly benefited from advancements in highly parallelised short-read sequencing methods, such as the

Illumina sequencing platform (Levy and Myers 2016). The cost, power and availability of these technologies and the associated field of bioinformatics have improved dramatically, such that even population-scale genomic analyses are now possible (Park and Kim 2016).

Sequencing technology advances have allowed WGS to be combined with BSA. In such BSA sequencing (BSA-seq) experiments, pooled DNA from each bulk is sequenced, with the resulting output containing short sequences from each individual in the pool. The contigs are mapped to a reference genome and software is then used to call single nucleotide polymorphisms (SNPs) between the two bulks. Allele frequency for each bulk is estimated by taking the number of mapped reads with each of two possible nucleotides at a SNP. This number, as a proportion of the total number or reads mapped to that nucleotide position, is a proxy for allele frequency. With a high enough sequencing depth and large enough bulks, allele frequency estimates can be scanned across the genome, allowing accurate genetic mapping of segregating traits (Magwene *et al* 2011).

3.1.6 Statistical considerations of bulked segregant analysis

The mapped allele frequency differences between BSA pools is a good indicator of the location in the genome of loci linked to a trait of interest (Parts *et al* 2011). However, a higher detection power can be achieved by using a *G*-statistic-based test such as that developed by Magwene *et al* (2011). The calculation of their statistic is shown in detail in chapter 2. The raw allele frequency difference value has a slow rate of decay around a causal locus, thus making it difficult to refine identification of causal loci beyond a large interval. However, the decrease in the value of *G* around the same locus is expected to be more rapid, giving a more precise estimation of the location of a causal locus on a chromosome (**Figure 3.12**). This *G* calculation also benefits from the inclusion of more parameters, such as genome-wide sequencing coverage rather than simply the depth of coverage at a SNP.

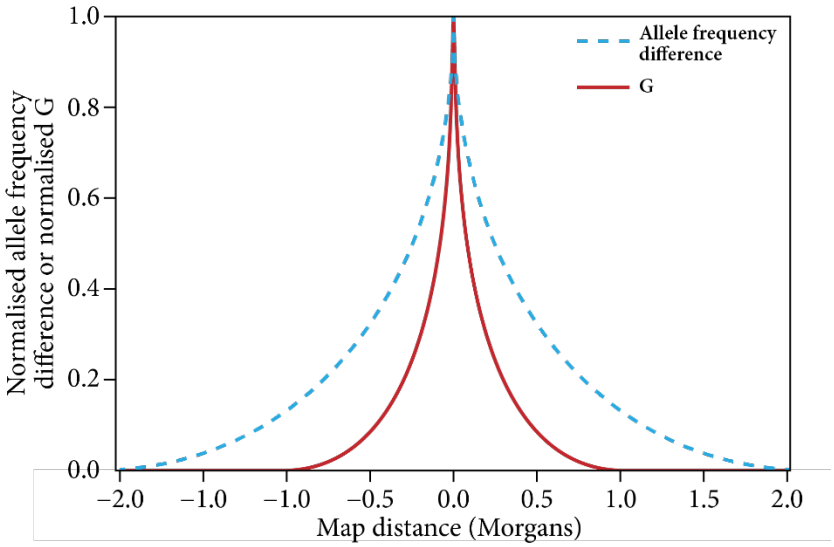


Figure 3.12 The allele frequency difference (dashed blue line) and the G statistic around a hypothetical causal SNP. Adapted from Magwene *et al* (2011).

Magwene *et al* (2011) also calculate a smoothed version of the *G* statistic, which they call *G*′. Calculating *G*′ involves averaging the *G* value across a sliding window of fixed width (typically 50 kb). They also use a smoothing kernel (tri-cube) to adjust neighbouring values and achieve a smoother distribution. Using *G*′ is helpful because the variation in the unprocessed *G* value prevents the detection of causal loci. This variation comes from two stages in the experimental process. Firstly, populations used in BSA studies segregate for more than one trait. This first source of variation can be minimised by using larger segregating populations, more individuals in each bulk and more highly inbred parents of the population. Secondly, the sequencing stage of BSA experiments introduces additional variation, which may come from differences in coverage across the genome, incomplete library preparation, misalignment of reads or problems with algorithms used to call SNPs. Protocol optimisation can help alleviate many of these concerns, but some stochasticity is likely to remain (Magwene *et al* 2011).

3.1.7 Fine-mapping by genotyping large populations

Once a phenotype has been mapped to a specific genomic region using BSA, it can be mapped more precisely using fine-mapping techniques. Every individual in a large population segregating for the trait of interest is phenotyped and genotyped at close intervals across the locus that was identified in the BSA results. Chance recombination events between markers will reveal which markers are most closely linked to the gene

responsible for the phenotype.

One method of doing this efficiently is to use allele-specific polymerase chain reaction (PCR). This method is described in detail in chapter 2. Briefly, the sequence of the 20-25 nucleotides immediately preceding a SNP are incorporated into two oligonucleotide primers: one ending with the nucleotide of the first allele at that SNP, and one ending with the nucleotide of the second allele. A common primer in the opposite direction is used to flank the other end of a short PCR product. Different fluorescent tags attached to each allele-specific primer allow identification of which primer has directed successful amplification in each reaction. In individuals homozygous at the SNP, there will only be one fluorescence peak. For heterozygous individuals, one would expect two peaks – one for each allele. This method has been commercialised by LGC Ltd (2013) in a technology called Kompetitive Allele-Specific PCR (KASP).

3.1.8 Aim of this work

My aim in the experiments detailed in this chapter is to determine the genetic basis of the white face phenotype in *A. m. pseudomajus*, a highly prevalent phenotype in the genus. This phenotype is part of a suite of patterns produced in *Antirrhinum* by accumulating magenta anthocyanin pigments in different parts of the flowers' petals. These other flower colour traits are each changed by one gene – *ROSEA* turns on anthocyanin production in the petals, *VENOSA* does the same in petal cells overlying veins in the dorsal petals and *ELUTA* restricts anthocyanin pigmentation to the centre of the flower. Given this, a hypothesis for the genetic basis of the white face phenotype is that an additional single gene of large effect is involved in regulating the white face phenotype in *A. m. pseudomajus*, thus adding to the suite of genes known to regulate magenta pigmentation. Perhaps an *A. majus* homologue of *Mimulus lewisii*'s *LAR1* might fulfil this role by producing competing colourless flavonols, or another transcription factor may switch off anthocyanin biosynthesis genes in a specified part of the flower in the same manner as *ELUTA*. According to these hypotheses, bulked segregant analysis would show a single peak either at a *LAR1* homologue or elsewhere in the genome when comparing pools of plants with and without the white face phenotype. Alternatively, white face may be a polygenic trait with several loci contributing to the final phenotype. Bulked segregant analysis would then show several distinct peaks when comparing white face and non-white face pools.

3.2 **Results:** Segregation of the white face phenotype in F2 populations

To identify the genetic basis of *A. m. pseudomajus*'s white face, I needed a population that segregated for the trait. Previous work on isolating genes from *A. m. pseudomajus* has involved crossing plants to the closely related *A. m. striatum* because the two subspecies are very similar both genetically and phenotypically, differing in little other than their flower colour. However, because *A. m. striatum* does not exhibit petal-wide magenta pigmentation, the effect of any gene restricting anthocyanin biosynthesis in the face of the flower would not be seen. Instead, *A. m. pseudomajus* was crossed to a cultivar of *A. majus* known as John Innes Stock 7 (JI7) (**Figure 3.13**). Plants in the JI7 line are darkly magenta-coloured throughout their petals without the white face phenotype seen in *A. m. pseudomajus*. This cultivar is also the one used for the assembly of the *A. majus* genome (Xue *et al*, in preparation).

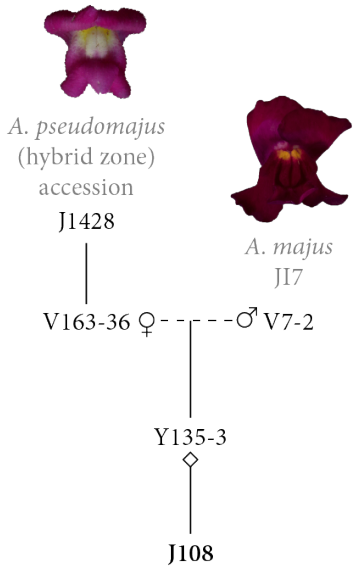


Figure 3.13 A pedigree of the plant family J108 grown for analysis of the white face phenotype. Seed from J1428 was collected near the village of Ventola, Ribes de Freser, Girona, Catalonia and grown at the John Innes Centre as V163. Plant 36 from that family was crossed to the *A. majus* research line JI7 (individual V7-2). Plant 3 from the F1 generation (Y135) was selfed to generate the family J108. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

I used an F2 population segregating for this white face/non-white face phenotype generated by crossing an *A. m. pseudomajus* plant grown from wild-collected seed to a JI7 plant. **Figure 3.14** shows the magenta flower colour variation seen in the progeny of this cross, ranging from the full pigmentation of the JI7 line to the strong white face of *A. m. pseudomajus*. In the first F2 population of 240 plants, family J108, 57 plants showed a strong white face phenotype, 125 had a weak white face and 58 had no white face. The statistical difference between this ratio and a 1:2:1 ratio expected for a trait caused by a semidominant allele at a single locus is not significant (p = 0.404). This suggests that *A. m. pseudomajus* carries a semidominant allele at a single locus that regulates the white face phenotype. The same F2 family also showed variation in other floral traits that were not considered, including in flower size and shape, and in the amount of yellow colouration at the foci, although these did not appear to be linked to the white face phenotype.

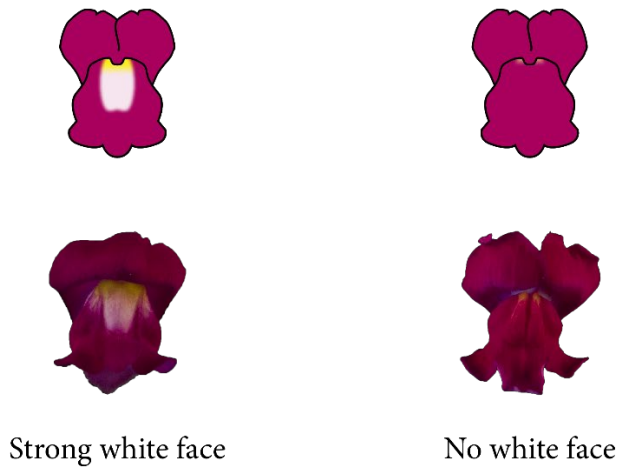


Figure 3.14 Magenta variation seen in the progeny of crosses between *A. m. pseudomajus* and the *A. majus* JI7 line. In some individuals, a large white patch is seen in the face region ('strong white face'); in others, the petals are magenta throughout ('no white face'). A schematic illustration (top) and example photograph (bottom) is shown for each version of the phenotype.

3.3 **Results:** Bulked segregant analysis of the white face trait from *A. m. pseudomajus*

Two pools of DNA were sequenced from the F2 population J108 segregating for the white face phenotype. The first comprised 13 plants showing the clearest strong white face phenotype, and the second comprised 29 plants showing the clearest full magenta

phenotype.

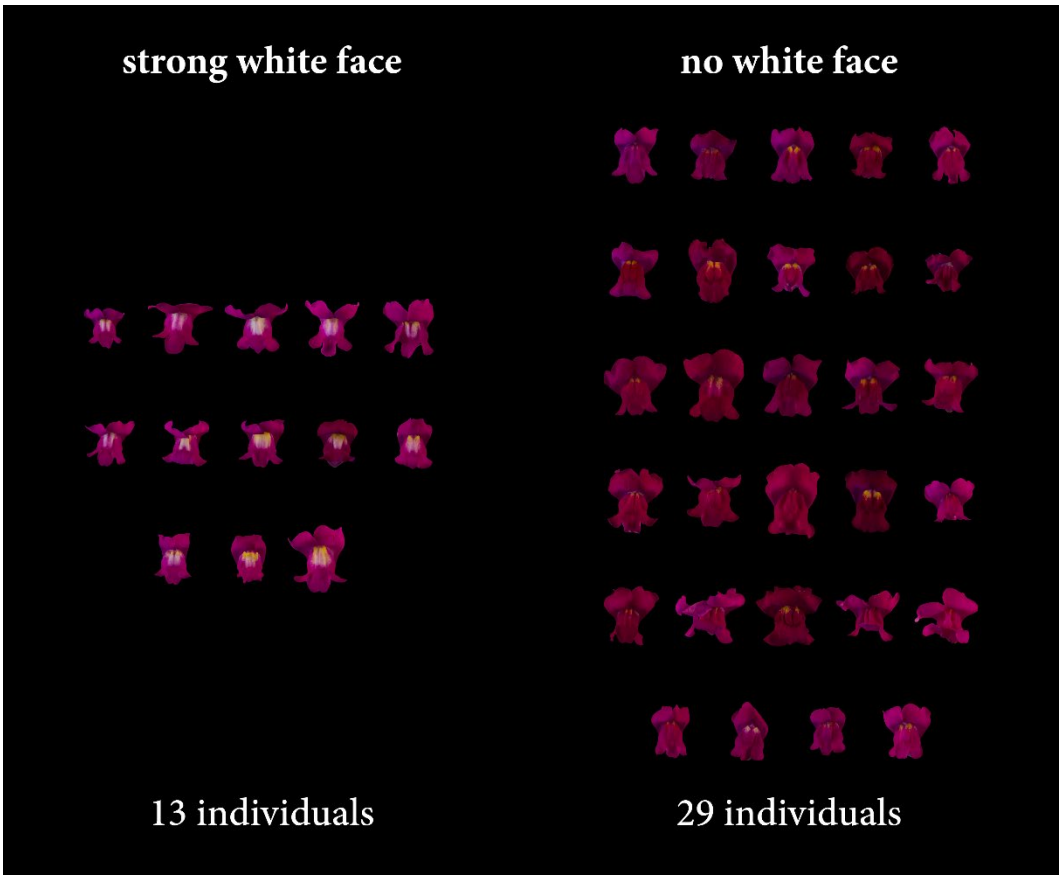


Figure 3.15 Flower photographs from the 42 individuals from J108 used to construct the bulked segregant analysis pools. DNA from the leaves of the 13 individuals whose photographs are pictured on the left was pooled to form the 'strong white face' pool. DNA from the leaves of the other 29 individuals were used to form the 'no white face' pool.

I identified SNPs in the resulting data relative to the JI7 reference genome and calculated the frequencies of the reference and non-reference alleles at each of these sites and the difference in allele frequency between the pools at each SNP (Δ SNP-index). From these allele frequencies, I calculated a G value at each site, testing the null hypothesis that there is no QTL linked to that site.

The bulked segregant analysis of J108 showed seven peaks that crossed the false discovery rate threshold, which was set at a q value of 0.01 (**Figure 3.16**). The main plot of interest is the middle of the three, showing G' values for SNPs averaged across 50 kb windows. Three of the peaks seen are known to be on loci misassembled in the current version of the *Antirrhinum* genome. These are the very narrow peaks on chromosomes 3, 4 and 5. They are shown in light grey on the graphs in **Figure 3.16**. G' values at these loci consistently match those seen on parts of chromosome 2,

suggesting that they are linked and that, in future versions of the genome, these loci should be assembled on chromosome 2. However, given that their exact positions on chromosome 2 is unknown, I have left them in their currently assembled positions for this work. There is also a section of chromosome 2 that is assembled in the incorrect orientation – this has been given a green background in this figure. Because of this, chromosome 2 should be treated as one peak rather than two.

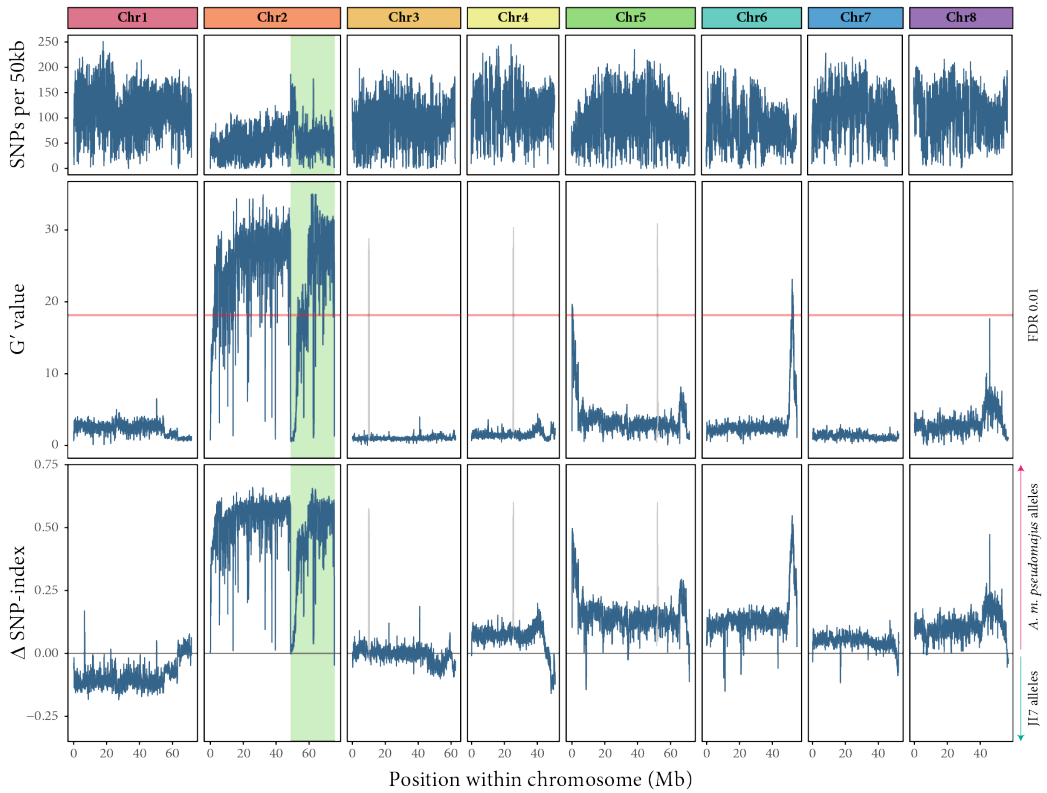


Figure 3.16 Bulked segregant analysis Manhattan plots for phenotypic extremes from family J108 segregating for the white face phenotype. The top plot shows the number of SNPs in each 50 kb window across each chromosome. The middle plot shows the *G′* value for each SNP. This is a version of the *G* value averaged across 50 kb windows and smoothed using a tri-cube kernel function. The red line on this plot represents a *G′* threshold corresponding to a false discovery rate of 0.01. The bottom plot shows the difference in allele frequency of each pool, again averaged and smoothed across 50 kb windows. A negative value indicates that, in a majority of sequencing reads, the JI7 reference genome nucleotide is found at SNPs in that window; a positive value indicates that a majority of reads have a non-reference nucleotide at SNPs in the window. Mismapped regions are shown in light grey, and a green background represents the section of chromosome 2 that is assembled in the wrong orientation.

Discarding the three misassembled regions and considering chromosome 2 as having

just one peak leaves three peaks to be investigated. Firstly, chromosome 2 shows very high G' values, with nearly the entire chromosome appearing as a peak. Secondly, chromosome 5 has a peak covering its first 3Mb. And thirdly, a large peak is found on chromosome 6, seen towards the end of the chromosome. Elevations in G' value can also be seen around the final 20Mb of chromosome 8 and the last 6Mb of chromosome 5, but these are lower than the false discovery rate threshold. The observance of numerous peaks usually means that a phenotype is polygenic, regulated by genes at several distinct QTL. However, this contradicts the 1:2:1 ratio of white face phenotypic classes seen in J108. This suggests that the plants in the two pools may have differed in more than just their white face phenotypes, perhaps because of the small number of individuals used.

The chromosome 2 peak is the highest of the three. It includes three known flower colour genes. *FLAVIA* (*FLA*), which encodes chalcone glucosyltransferase (CGT), an enzyme involved in the biosynthesis of the yellow aurone pigment in *Antirrhinum* (Ono *et al* 2006, Boell *et al* unpublished results) (**Figure 3.17**). The other two known flower colour genes on the chromosome are *CHALCONE ISOMERASE* (*CHI*), involved in the early stages of anthocyanin biosynthesis and *AUREUSIDIN SYNTHASE1* (*AS1*), which works alongside CGT to make aurone.

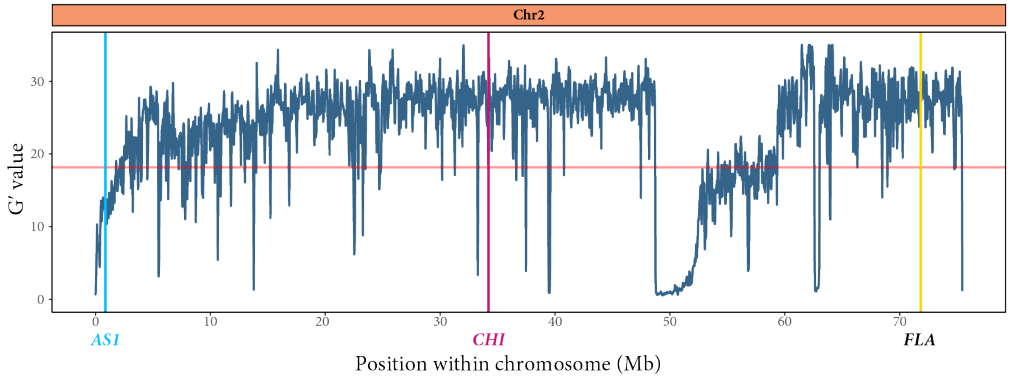


Figure 3.17 A closeup view of chromosome 2 showing G' values across that chromosome (dark blue line), using the same BSA data as **Figure 3.16**. The positions of three genes found on this chromosome are shown with vertical lines: *AUREUSIDIN SYNTHASE 1* (light blue), *CHALCONE ISOMERASE* (magenta) and *FLAVIA* (gold). The pale red horizontal line represents a G' threshold corresponding to a false discovery rate of 0.01.

The anthocyanin biosynthetic gene *CHI* itself may be contributing to the white face phenotype or to variation in magenta. The chalcone isomerase enzyme it encodes is involved in the early stages of the flavonoid biosynthetic pathway. A change in the

function of the chalcone isomerase protein is unlikely, given that anthocyanins still accumulate in the flowers outside the face region in plants with a strong white face phenotype. However, a change in the *cis*-regulatory region of this gene could alter the regulation of the gene without compromising the function of the chalcone isomerase enzyme. This would mean a marked difference in the way the magenta pigment is regulated in the face region compared with other parts of the flower. Interspecies differences in patterns produced by anthocyanins have previously only mapped to transcription factors – no differences have been found at the loci encoding biosynthetic genes. Furthermore, the transcription factors described – *ROS*, *EL* and *VE* – interact with stages in the anthocyanin pathway much later than where chalcone isomerase acts (Schwinn *et al* 2006).

Another possible explanation for the peak on chromosome 2 is that the individuals compared in this analysis differed not only for magenta pigmentation, but for yellow pigmentation too. In J108, the size of the yellow region seen varied from a tiny hint at the foci to a much greater degree of yellow pigmentation covering much of the flower face. I looked back at the photographs taken of the flowers of J108, and in particular at those individuals used to construct the BSA pools. There was a great deal of variation in the amount of yellow in the face region of these flowers; some flowers had very weak and restricted yellow pigmentation, whereas others had strong yellow pigmentation that spread much further down the face (**Figure 3.18**). Such variation in the amount and intensity of yellow pigmentation may be explained by segregation of alleles at the *FLA* locus; the *FLA* allele of JI7 is known to produce a much stronger yellow colour than that of *A. m. pseudomajus* (Boell *et al* unpublished results). The lack of yellow may make the white face phenotype more visible when scoring; conversely, rich yellow colouration may make the magenta colour of non-white face individuals appear darker. Because the two bulks were the tail ends of the phenotypic distribution between white face and non-white face without consideration for other traits, some unconscious bias with regard to yellow may have influenced the construction of the pools.

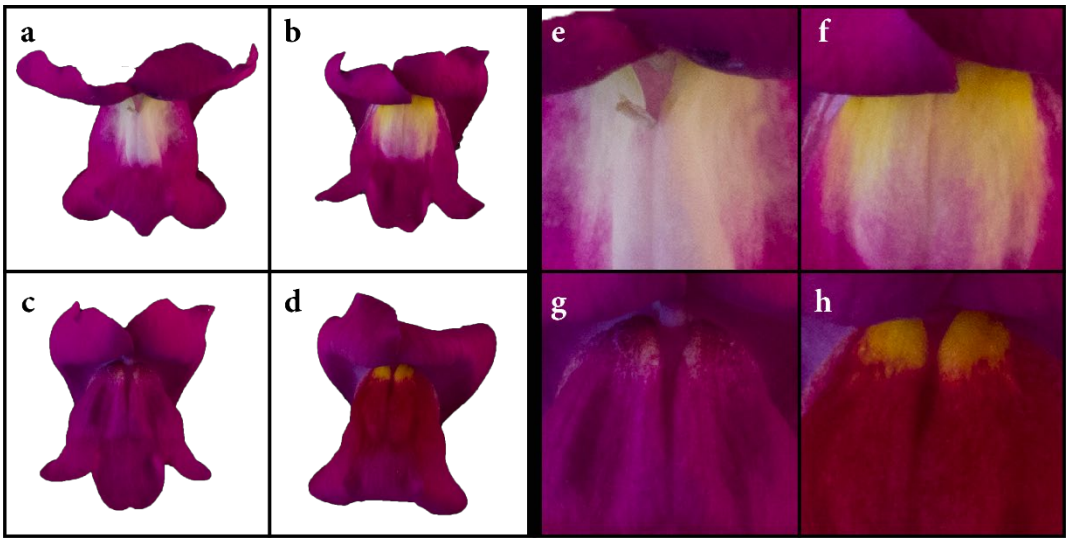


Figure 3.18 Yellow variation seen in J108. In some individuals, the face of the flower only had a hint of yellow pigmentation (**a** and **c**), whereas others had much stronger yellow (**b** and **d**). The same flower photographs are shown in **e** to **h**, with their face regions magnified.

To determine whether the signal on chromosome 2 was linked to the white face phenotype or to the variation seen in yellow pigmentation, the individuals used for the two J108 BSA pools were genotyped using KASP markers (LGC Ltd 2013). These markers determine the presence or absence of two alleles at a SNP for each individual tested. If only one is present, the individual is homozygous for that allele; if both are detected, the individual is heterozygous at the SNP of interest. In this analysis, the marker used was in the promoter region of *FLA* and the assay tested for the JI7 and *A. m. pseudomajus* alleles. In the white face pool, nine plants were homozygous for the *A. m. pseudomajus* allele of *FLA*, and the other four were heterozygous. In the non-white face pool, five plants were homozygous for the *A. m. pseudomajus* allele, 18 were homozygous for the JI7 allele and six were heterozygous. This means that in the white face pool, there were 22 *A. m. pseudomajus FLA* alleles and only four JI7 *FLA* alleles. In the non-white face pool, however, there were 42 JI7 *FLA* alleles but only 16 *A. m. pseudomajus FLA* alleles. This predicts a peak in allele frequency difference at *FLA* and, therefore, at least some of the G' value peak on chromosome 2 is because of this unconscious bias.

The width of the peak on chromosome 2 also requires explaining. In work attempting to genetically map traits from *Antirrhinum* species using the JI7 line, a lack of

recombination on chromosome 2 is consistently observed (Boell *et al* unpublished results). **Figure 3.19** shows the results of four bulked segregant analyses from chapter 4 of this thesis, in which the pools tested were also from F2 populations derived from crosses between *Antirrhinum* species and JI7. Here too, all peaks on chromosome 2 are nearly chromosome-wide. The cause of the lack of recombination on this chromosome has not been determined, but such effects are often caused by inversions. Given that this effect is seen when analysing variation from many different *Antirrhinum* species, this would mean an inversion in the chromosome 2 of JI7, relative to other *Antirrhinum* accessions. Such a phenomenon may also explain the difficulty in correctly assembling chromosome 2. Issues with this chromosome's assembly are discussed in section 4.6.4.

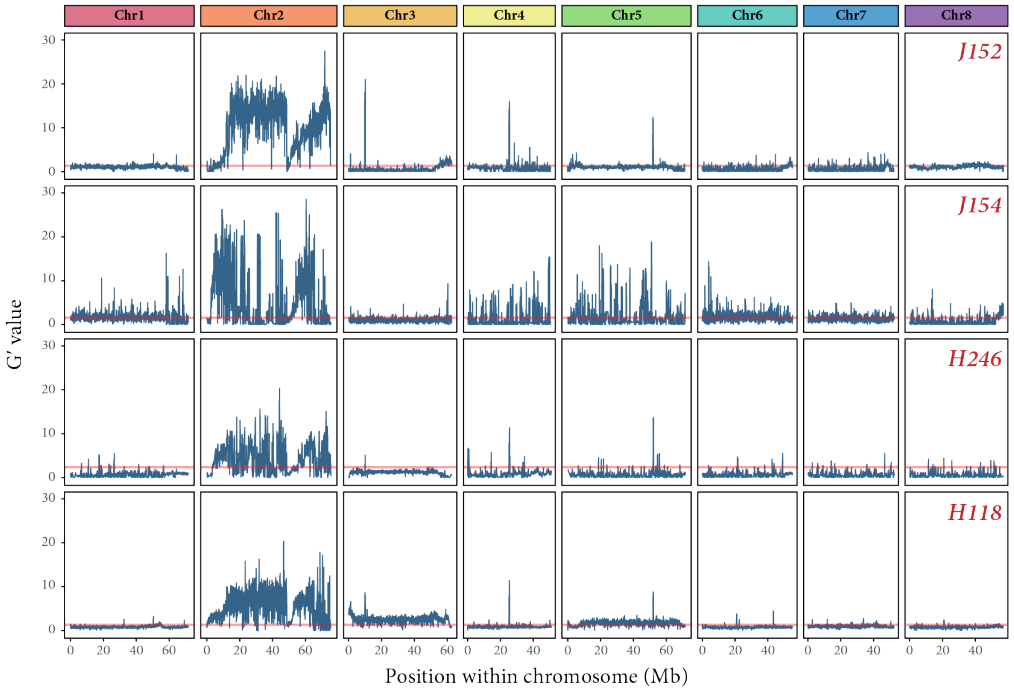


Figure 3.19 Manhattan plots showing the result of bulked segregant analyses described in chapter 4 of this thesis, which focuses on yellow colour variation. As is the case in family J108 (**Figure 3.16**), peaks on chromosome 2 in these analyses cover nearly the whole chromosome. The blue line shows the G' value for each SNP, a variation on the G statistic averaged across 50 kb windows and smoothed using a tri-cube kernel function. Red horizontal lines on each plot represent a false discovery rate of 0.01.

The peak on chromosome 5 (**Figure 3.16** and enlarged in **Figure 3.20**) is at the very start of the chromosome, with the first 3Mb of the chromosome showing an elevated G' value between 5 and 10. A lower elevation in G' ($G' < 5$) is seen around the 66Mb position on the chromosome, although this peak does not cross the false discovery rate

threshold (a q value of 0.01). Neither of these two regions contain genes previously known to be involved in flower colour in *Antirrhinum*. The start of chromosome 5, therefore, is of particular interest as the height of its peak is consistent with it containing a white face-linked causal locus. A previously unidentified gene involved in pollinator-attracting flower colour patterns may therefore be found in this area. Alternatively, similarly to chromosome 2, it may be a result of the pooling strategy, containing loci regulating a trait that was inadvertently selected for in the pools.

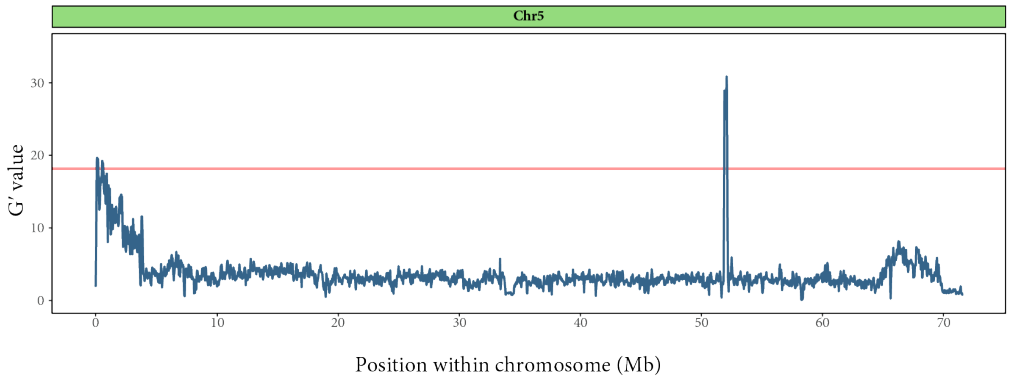


Figure 3.20 A closeup view of chromosome 5 showing G' values across that chromosome (dark blue line), using the same BSA data as **Figure 3.16**. The pale red horizontal line represents a G' threshold corresponding to a false discovery rate of 0.01.

I extracted the sequences for the first 3Mb and final 10Mb of chromosome 5 to look for sequences homologous to genes in other plant species. I performed a blastx analysis using the basic local alignment search tool (BLAST) (Altschul *et al* 1990). The blastx tool compares the amino acid translation of a query sequence against a large database of protein sequences and is designed to find sequences that are homologous to protein-coding genes. I used the NCBI non-redundant database for this search (NCBI Resource Coordinators 2017). None of the genes found in this analysis appear to be involved in flower colour or, more generally, in the regulation or synthesis of flavonoids. There are also no MYB-like transcription factors, which are typical regulators of magenta flower colour in *Antirrhinum*, encoded within these two regions. If flower colour-regulating genes are located in these regions, therefore, they would be new genes without homologues described in the NCBI database, and would likely involve a mechanism other than MYB-like transcription factors. However, it is possible that the peaks may be linked to phenotypes that were not selected for in the pools, but that happened to differ between some of the plants compared. Such a difference between only some of the individuals in the bulks would also explain these

peaks being low compared to some of the other peaks.

The peak on chromosome 6 (**Figure 3.21**) contains the linked *ROSEA* (*ROS*) and *ELUTA* (*EL*) genes, although it is not possible with this data to determine whether one of these genes is more closely linked to the phenotype than the other. That either of these genes are linked to the white face phenotype is surprising. Both are known to regulate magenta pigmentation, but the effects of both have been previously described: *ROS* activates late-pathway anthocyanin biosynthetic genes throughout the petal lobes, and *EL* restricts the activity of the same genes outside the centre of the flower. Both are also thought to have the same allelic configuration in *A. m. pseudomajus* as in *A. majus* var. JI7 – dominant *ROS* (*ROS*/-) and recessive *EL* (*el*/*el*). These results, however, imply that a difference between *A. m. pseudomajus* and JI7 at or near the *ROS-EL* locus may be partly responsible for the difference in their magenta phenotypes. Alternatively, this peak could represent differences in magenta intensity that make the white face easier to score, without being involved in suppressing anthocyanin production in the face directly.

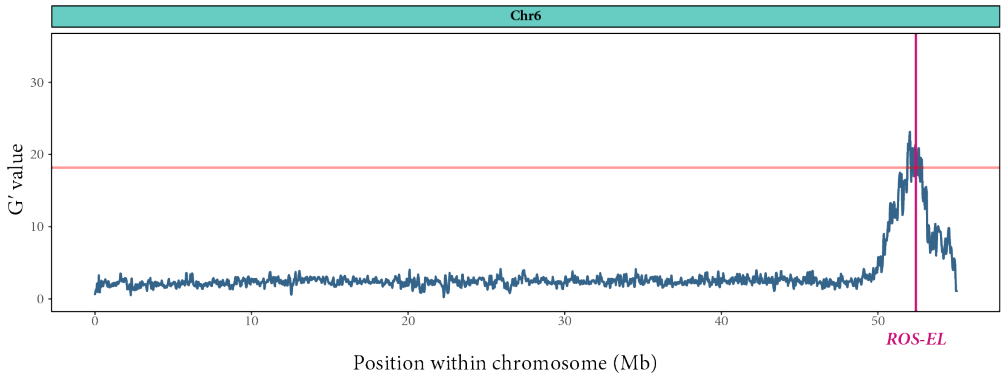


Figure 3.21 A closeup view of chromosome 6 showing *G′* values across that chromosome (dark blue line), using the same BSA data as **Figure 3.16**. The position of the *ROS-EL* locus is shown with a magenta vertical line. The pale red horizontal line represents a *G′* threshold corresponding to a false discovery rate of 0.01.

The low peak on chromosome 8 is highest around the 45Mb position along the chromosome (**Figure 3.22**). This region does not contain genes previously known to be involved in flower colour. *VENOSA*, a gene encoding a MYB-like transcription factor that regulates anthocyanin pigmentation in the dorsal petals, is located on this chromosome, but at the start of the chromosome, 43-45Mb away from the peak. As with the results from chromosome 5, I used a blastx search to look for regions homologous to protein-coding genes from other species at this peak, but did not find

any likely to be involved in flower colour regulation.

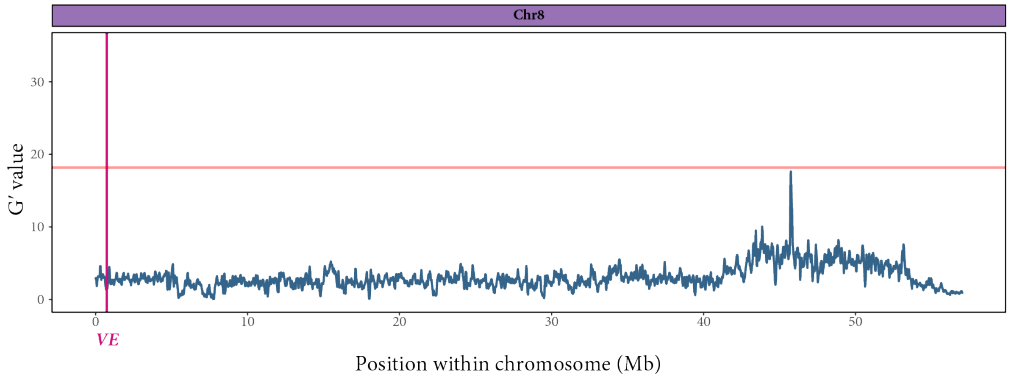


Figure 3.22 A closeup view of chromosome 8 showing G' values across that chromosome (dark blue line), using the same BSA data as **Figure 3.16**. The position of the *VE* locus is shown with a magenta vertical line. The pale red horizontal line represents a G' threshold corresponding to a false discovery rate of 0.01.

None of the peaks described above contain a homologue to the *LAR1* gene in *Mimulus lewisii* described by Yuan *et al* (2016). The *A. majus LAR1* homologue is located on chromosome 1, but no peak in allele frequency difference between plants with and without the white face phenotype can be seen on this chromosome in **Figure 3.16**. This suggests that the white face phenotype is not analogous to the light area seen in *M. lewisii*.

The Δ-SNP index plot on the bottom row in **Figure 3.16** is based on the work of Takagi *et al* (2013), who developed a method for calculating the allele frequency differences at individual SNPs between bulks with extreme phenotypes in segregating populations. First, the proportion of reads mapped to a SNP that carry the non-reference genome allele is calculated in each bulk. Then, the value for the 'wild type' pool (non-white face here) is subtracted from the value of the 'novel' phenotype pool (strong white face). This gives the Δ-SNP index for that SNP, and the same calculation is performed for all SNPs in the genome. For this analysis, I applied the same sliding window averaging (50 kb) and smoothing kernel (tri-cube) as I did the G' value so that the effect of outliers would be minimised, giving a cleaner signal. Full details of this method are given in chapter 2. Genomic positions not linked to a locus associated with the trait of interest (strong white face in this case) should have a Δ-SNP index value of zero, because 50% of alleles in both pools would come from each parent. If a genomic position has a positive value on this graph, it means that the white face pool held a

higher proportion of *A. m. pseudomajus* alleles than the non-white face pool. Conversely, a negative value indicates that the white face pool held a lower proportion of non-reference alleles than its non-white face counterpart. The peaks on chromosomes 2, 5, 6 and 8 all have positive values, which means that loci within these peaks are enriched for the *A. m. pseudomajus* allele in the white face pool compared to the non-white face pool.

The baseline Δ-SNP index for most of the genome does not lie at 0 in **Figure 3.16** – most genomic positions apart from most of chromosome 3 have a positive or negative value. In fact, most genomic positions have a higher proportion of *A. m. pseudomajus* alleles in one pool compared to the other. Only on chromosome 1 does the white face pool contain a higher proportion of reference (JI7) alleles than the non-white face pool, whereas for chromosome 2 and chromosomes 4 to 8 the reverse is true. This could be a sign that loci on these chromosomes may be enhancing the white face phenotype. Alternatively, this may be noise as a result of each pool containing relatively few individuals and thus differing by chance at some locations within the genome. A third possible reason is that plants were inadvertently selected to be more like *A. m. pseudomajus* in the white face pool and more like JI7 in the non-white face pool. Such differences may include flower shape and size, which may show subtle differences between the parents.

3.4 **Results:** Genotyping a larger population for *ROS-EL* to look at the relationship between genotype and phenotype

In order to map the white face phenotype at a finer scale and to determine whether loci at the *ROS-EL* locus or chromosome 2 were linked to the white face phenotype, I designed a series of SNP markers based on my BSA results (Section 8.1, Appendix 1). Using these KASP markers, I scored and genotyped a new F2 population of 500 plants grown in 2016. This population, L124, was grown from selfed seed from the same parent as J108, but from a separate capsule (**Figure 3.23**).

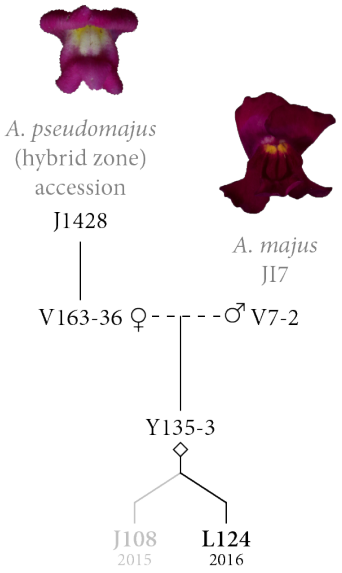


Figure 3.23 A pedigree of the plant family grown for analysis of the white face phenotype. Seed from J1428 was collected near the village of Ventola, Ribes de Freser, Girona, Catalonia and grown at the John Innes Centre as V163. Plant 36 from that family was crossed to the *A. majus* research line JI7 (individual V7-2). Plant 3 from the F1 generation (Y135) was selfed to generate the family L124. This was the same F1 plant that gave rise to J108. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

I also developed a more precise scoring system for the white face phenotype, giving each plant a magenta score between 0 (no white face) and 5 (strong white face) (**Figure 3.24**). To help with this phenotyping, I used an approximation of a multispectral imaging technique to visualise the flowers' colours more accurately (described in detail in chapter 2). In doing so, I hoped to distinguish as much as possible between the yellow and magenta pigmentation of each flower. To achieve this, I converted each image to black and white while applying a colour filter. As a result, in the images converted to visualise magenta, areas with the most intense magenta colouration appears dark grey or black, and areas with the least magenta colouration appears light grey or white, regardless of the presence of other colours. Similarly, yellow-coloured parts of the flower are the darkest in the yellow-converted images. I assigned a score between 0 and 5 based on visual inspection of the magenta-converted images. **Figure 3.24** shows the original images of representative flowers given each white face score, along with the colour-converted images. This scoring system is not perfect, as I did not

see any flowers with a magenta score of 0 but with strong yellow pigmentation too, suggesting that accumulation of aurones in the foci prevents anthocyanin pigmentation at the very top of the flower face.

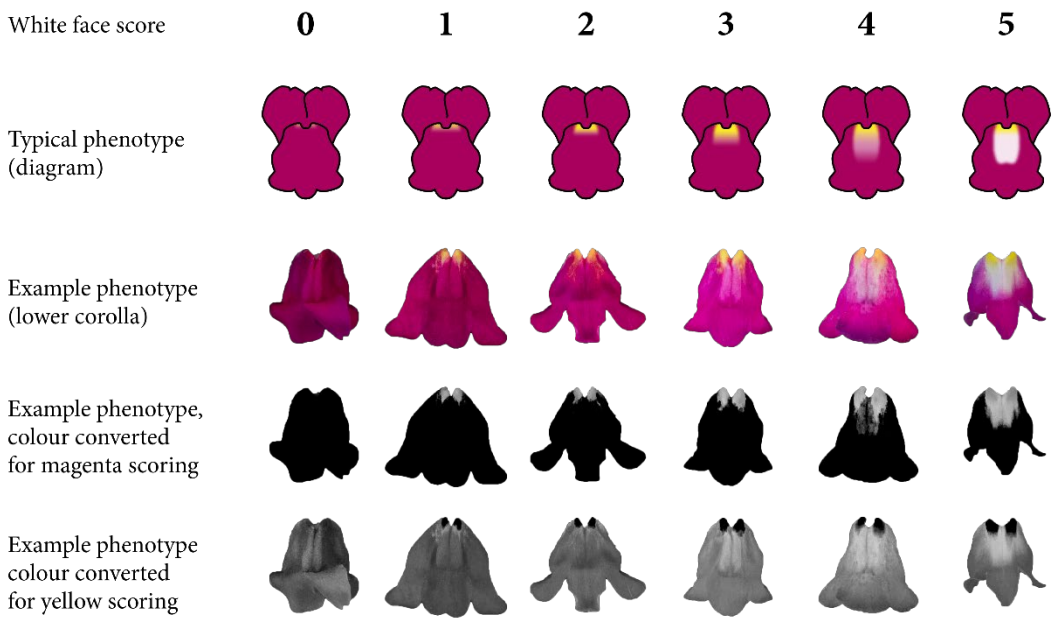


Figure 3.24 Revised flower colour scoring system used for phenotyping plants for the white face trait. The scores assigned range from full magenta pigmentation with no white (0) to a large white patch covering the face of the flower (5). The upper row shows a schematic representation of each score. The second row shows a representative photograph of a flower given each score, with the dorsal petals removed so as not to obscure the flowers' faces. The third row shows the same flower photographs seen in the second row, but converted to black and white with magenta pixels darkened. The same colour conversion has been applied in the fourth row, but with yellow pixels darkened instead of magenta ones.

As with J108, the majority of L124 plants showed a weak white face phenotype. Of the 498 plants, 42 plants (8.4%) had a white face score of 0 or 1 (no white face), 304 plants (61.0%) had a score of 2 or 3 (weak white face) and 152 plants (30%) had a score of 4 or 5 (strong white face). The number of plants given each score is shown in full in **Figure 3.25**. This ratio differs significantly from the 1:2:1 ratio seen in J108, with a *G* test for goodness of fit giving $p < 0.01$.

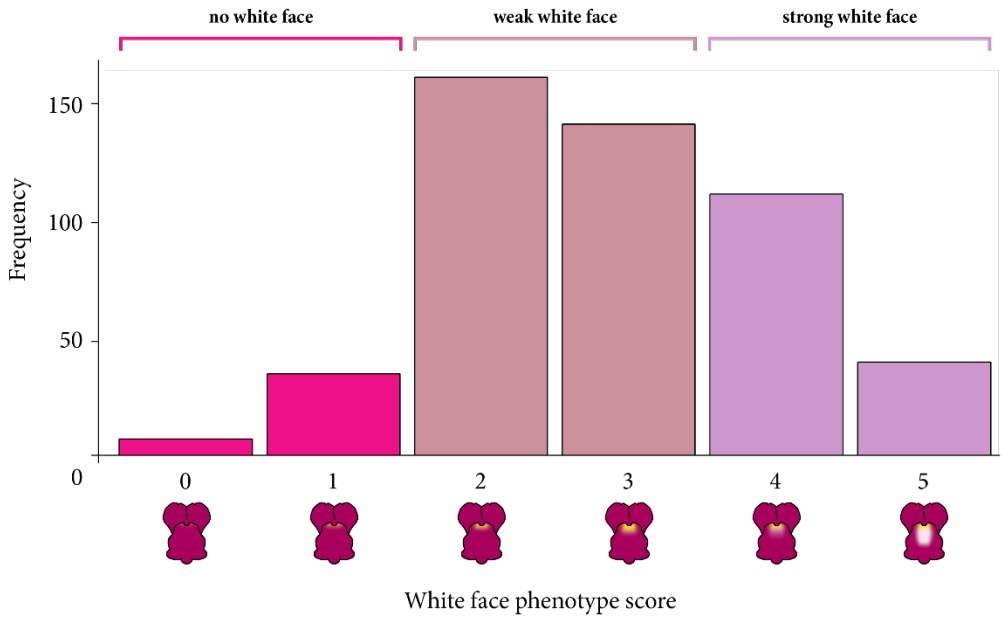


Figure 3.25 Frequency of each white face phenotype score in L124 (n = 498), grouped according to phenotype description. The diagrams below the bars show a typical visual representation of each score.

I designed and tested 25 sets of primers at and near the *ROS-EL* locus. These are shown in detail in **Section 8.1** (Appendix). I used one KASP marker in each gene for this work: one marker in the second intron of *ROS1* (52,319,793 bp along chromosome 5) and one in the exon of *EL* (52,491,807 bp along chromosome 5). The white face phenotype was closely associated with the genotypes at the *ROS-EL* locus. This association is shown in **Figure 3.26**, where I show the proportions of individuals with each *ROS* and *EL* genotype, sorted by white face phenotype. In summary, of the plants showing the strongest white face phenotype (score 5), 77% were homozygous for the *A. m. pseudomajus* allele of *ROS* (genotype ROS^{P}/ROS^{P}) and 65% were homozygous for the *A. m. pseudomajus* allele of *EL* (genotype EL^{P}/EL^{P}); the exceptions were heterozygous (ROS^{P}/ROS^{7} and EL^{P}/EL^{7}, where a superscript *7* indicates the *A. majus* stock 7 allele). Of those showing the clearest non-white face phenotype (score 0), 71% had the genotype ROS^{7}/ROS^{7} and 57% had the genotype EL^{7}/EL^{7}; exceptions, again, were heterozygous (ROS^{P}/ROS^{7} and EL^{P}/EL^{7}). Looking at the two extreme phenotypes (scores 0 and 5), around 75% of the variation in white face phenotype is accounted for by *ROS-EL*. There are three possibilities that could explain the residual variation in this phenotype. First, a second gene, unlinked to *ROS-EL*, may act as a modifier of the white face phenotype. Second, there may be variation in the white face phenotype that is not controlled genetically – environmental differences might alter

the amount of magenta pigmentation on the flower face. And third, some of the genotyping results may be inaccurate, due either to problems with the oligonucleotide markers or sample contamination.

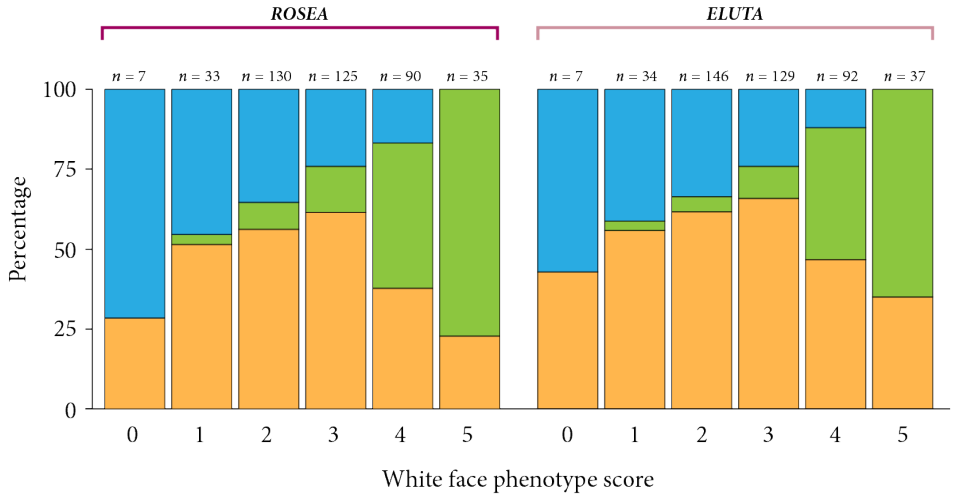


Figure 3.26 Percentages of each of the three *ROS* and *EL* genotypes present in individuals from the family L124 grouped by white face phenotype score. Individuals homozygous for the JI7 allele are coloured in blue, those homozygous for the *A. m. pseudomajus* allele are coloured in green and heterozygous individuals are shown in yellow. The number of individuals falling within each group is shown above its bar. These are not the same for both genes because genotypes could not confidently be called for all reactions.

I also recorded the yellow flower phenotypes of plants in L124 based on the yellow-converted images (**Figure 3.24**) and genotyped them to determine whether differences at the *FLA* locus can explain the variation in the amount of yellow seen on the flowers. As shown in **Figure 3.27**, most plants (296 out of 498) had yellow foci on their flowers. An additional 104 had spread yellow pigmentation, where the yellow on the flower face extended down from the foci. The remaining plants had little (57 plants) or no yellow (41 plants) on their flowers. I performed *G* tests for goodness of fit to determine whether the yellow phenotypic ratios in this family matched those expected for a single gene. The ratio between the number of flowers with spread yellow, restricted yellow and little-to-no yellow differed significantly ($p < 0.01$) from the 1:2:1 ratio that would be expected if the variation was governed by a single locus with a semidominant allele. The ratio between the number of flowers with spread yellow and restricted/suppressed yellow also differed significantly ($p < 0.01$) from the 3:1 ratio that would be expected if a single causal locus had a fully dominant spread-yellow allele. This could mean that yellow variation in this family is controlled by more than one gene.

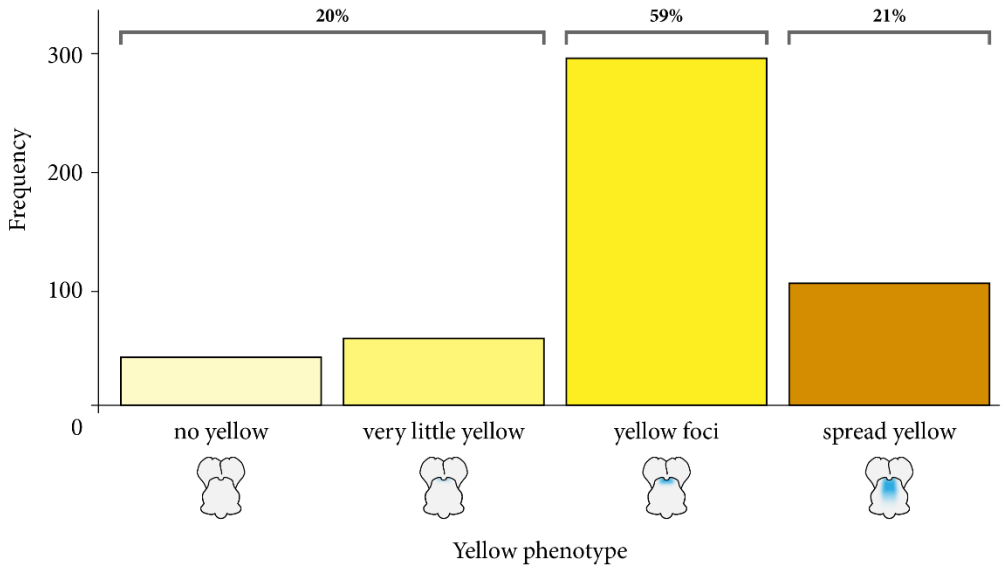


Figure 3.27 Frequency of each yellow phenotype in L124 (n = 498), grouped according to phenotype description. The diagrams below the bars show a typical visual representation of each score, but using blue to illustrate yellow pigmentation to make it easier to see.

The *FLA* genotypes for this family showed the expected 1:2:1 ratio between individuals carrying only the *A. m. pseudomajus* allele, both alleles and only the JI7 allele, respectively. Some association could be found between *FLA* genotype and yellow phenotype (**Figure 3.28**). Plants with little or no yellow in their flowers were mostly homozygous for the *A. m. pseudomajus* allele of *FLA* (80%), with some heterozygotes (20%). Plants scored as having yellow foci not spreading down the flower face could have any of the three *FLA* genotypes, with 58% heterozygotes, 30% JI7 homozygotes and 12% *A. m. pseudomajus* homozygotes. Similarly, in the group of plants with spread yellow on the faces of their flowers, 56% were heterozygous and 44% were homozygous for the JI7 allele; none of these plants were homozygous for the *A. m. pseudomajus* allele. These results suggest that the *A. m. pseudomajus* allele of *FLA* generally confers less yellow than that of JI7, but that another gene may also regulate the restriction of yellow in L124.

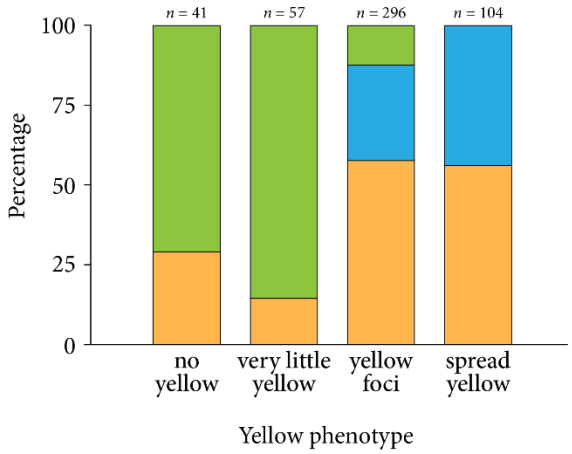


Figure 3.28 Percentages of each of the three *FLA* genotypes present in individuals from L124 grouped by yellow phenotype. Individuals homozygous for the JI7 allele are coloured in blue, those homozygous for the *A. m. pseudomajus* allele are coloured in green, and heterozygous individuals are shown in yellow. The number of individuals falling within each group is shown above its bar.

3.5 **Results:** A second round of bulked segregant analysis shows that just one peak is linked to the white face phenotype

The incomplete linkage between the white face phenotype and the genotype at *ROS* and *EL* suggested that an additional unlinked locus may affect the intensity of the phenotype. Alternatively, the results may indicate sample contamination or inaccurate genotyping. I generated additional pools for sequencing to attempt to determine whether an additional contributing locus exists and identify it, this time using plants from L124 that I had genotyped for *ROS* and *EL* and scored more accurately. I grouped the magenta-converted flower images according to *ROS-EL* genotype and selected the white face/non-white face phenotypic extremes. I sequenced four pools for this population, each comprising DNA from 20 plants:

1. **Strong white face, *A. m. pseudomajus* alleles of *ROS* and *EL***. These were the individuals that had the strongest white face phenotypes and were confirmed to be homozygous for the *A. m. pseudomajus* alleles of both *ROS* and *EL*. These all had magenta scores of 5.
2. **No/very little white face, *A. m. pseudomajus* alleles of *ROS* and *EL***. These were the ROS^{P}-EL^{P} homozygotes that had the fullest magenta colouration with the least amount of white face. These had magenta scores of 1 and 2.
3. **Weak white face, JI7 alleles of *ROS* and *EL***. These were the ROS^{7}-EL^{7}

homozygotes that had the strongest white face phenotypes in that group. These had magenta scores of 3 and 4.

4. **No white face, JI7 alleles of *ROS* and *EL***. In the fourth pool, all the plants had a non-white face phenotype (score 0) and were homozygous for the JI7 alleles of *ROS* and *EL*.

I then performed bulked segregant analyses on six pairwise comparisons. I used these analyses to test two alternate hypotheses. Hypothesis 1 is a null hypothesis and states that the *ROS-EL* locus is solely responsible for the white face phenotype and that no other locus is linked to this trait; any additional variation is due to environmental effects. Hypothesis 2 states that one or more loci in addition to *ROS-EL* are differentially fixed between plants with opposing phenotype. The comparisons I used to test these hypotheses, and the expected outcome in each case, are shown in **Table 3.1**.

Table 3.1 Summary of the six pairwise comparisons I used to compare the sequenced pools in my second round of bulked segregant analysis. The final two columns show the expected results if hypothesis 1 (only *ROS-EL* is involved in changing the white face phenotype) or hypothesis 2 (*ROS-EL* and one or more additional loci are collectively involved in changing the white face phenotype) were true.

	Pools compared	Summary of comparison	Hypothesis 1 result	Hypothesis 2 result
A	Pool 1 by pool 2	Different phenotypes, but same *ROS-EL* genotype	No peaks	One or more peaks, but not at *ROS-EL*
B	Pool 1 by pool 3	Same phenotype, different *ROS-EL* genotypes	Peak at *ROS-EL* only	Peak at *ROS-EL* only
C	Pool 1 by pool 4	Different phenotypes, different *ROS-EL* genotypes	Peak at *ROS-EL* only	Peak at *ROS-EL* and at one or more other loci
D	Pool 2 by pool 3	Different phenotypes, different *ROS-EL* genotypes	Peak at *ROS-EL* only	Peak at *ROS-EL* and at one or more other loci
E	Pool 2 by pool 4	Same phenotype, different *ROS-EL* genotypes	Peak at *ROS-EL* only	Peak at *ROS-EL* only
F	Pool 3 by pool 4	Different phenotypes, but same *ROS-EL* genotype	No peaks	One or more peaks, but not at *ROS-EL*

Results of these analyses (**Figure 3.29**) show that when two pools are analysed and have the same genotype at *ROS-EL*, there is no peak anywhere in the genome (**Figure 3.29 A** and **F**). When two pools analysed have different genotypes at *ROS-EL*, a peak is seen at the *ROS-EL* locus, but not elsewhere in the genome, suggesting that no loci on other chromosomes contribute to the white face phenotype. This is consistent with hypothesis 1 being correct.

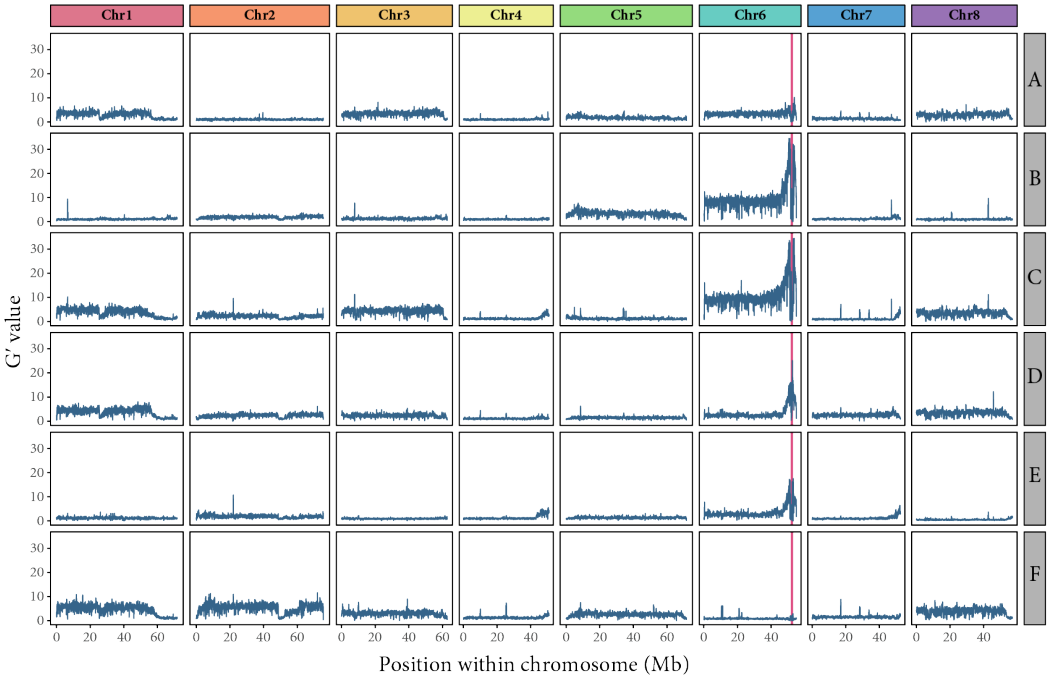


Figure 3.29 Manhattan plots showing the genome-wide *G′* values (*G* values averaged across a distance of 50 kb and adjusted using a tri-cube smoothing kernel) for six bulked segregant analyses. The names of the comparison made in each analysis is shown in the grey box to the right of each row and correspond to the descriptions in **Table 3.1**. In summary, they are: (**A**) pool 1 × pool 2, with different phenotypes but the same *ROS-EL* genotype; (**B**) pool 1 × pool 3, with the same phenotype but different *ROS-EL* genotypes; (**C**) pool 1 × pool 4, with different phenotypes and different *ROS-EL* genotypes; (**D**) pool 2 × pool 3, with different phenotypes and different *ROS-EL* genotypes; (**E**) pool 2 × pool 4, with the same phenotype but different *ROS-EL* genotypes; and (**F**) pool 3 × pool 4, with different phenotypes but the same *ROS-EL* genotype. The plots in each row are arranged horizontally by chromosome. Details of the pools are as follows: in pool 1, all plants have a strong white face phenotype and are homozygous for the *A. m. pseudomajus* allele at the *ROS-EL* locus; in pool 2, all plants have a non-white face phenotype and are homozygous for the *A. m. pseudomajus* allele at the *ROS-EL* locus; in pool 3, all plants have a weak-to-strong white face phenotype and are homozygous for the *A. majus* var. JI7 allele at the *ROS-EL* locus; in pool 4, all plants have a non-white face phenotype and are homozygous for the *A. majus* var. JI7 allele at the *ROS-EL* locus. The pink line in each of the chromosome 6 plots shows the location of the *ROS-EL* locus.

The peak at *ROS-EL* in **Figure 3.29** is highest and widest in analyses B (pool 1 × pool 3, same phenotype but different *ROS-EL* genotypes) and C (pool 1 × pool 4, different phenotypes and different *ROS-EL* genotypes). Here, all of chromosome 6 shows some elevation in *G′*, with a further elevation near the *ROS-EL* locus itself. These two plots

differ markedly from those for analyses D (pool 2 × pool 3, different phenotypes and different *ROS-EL* genotypes) and E (pool 2 × pool 4, same phenotype but different *ROS-EL* genotypes). This could mean that an additional locus on chromosome 6 modifies the phenotype subtly, either by enhancing the white face phenotype in *A. m. pseudomajus* or by reducing the white face phenotype in *A. majus* var. JI7. However, there are no high peaks seen in either analysis A or F (different phenotypes but the same *ROS-EL* genotype). If there was a modifier acting on the white face phenotype, a peak would be expected in one of these plots. This suggests that there is no genetic modifier and that unexplained variation in L124 is either environmental or because of incorrect genotyping.

A closer look at the genotyping used as the basis of this round of bulked segregant analysis reveals some discrepancies, suggesting that I misgenotyped some of the plants. I performed the genotyping for L124 using KASP primers, which I designed to give an allele-specific fluorescence when PCR plates containing the PCR-amplified samples are read under ultraviolet light. The absolute intensity of the fluorescence is not itself a reliable measure of the amount of each PCR product present as different primers can amplify better than others. To overcome this, the software used to read these genotyping plates (Bio-Rad CFX Manager) uses an algorithm to cluster values together based on the fluorescence intensity at each wavelength in each well. But re-genotyping these 80 plants showed that eight of the 20 individuals in pool 2, which should be homozygous for the *A. m. pseudomajus* allele of *ROS-EL*, were in fact heterozygous, and therefore also carried the *A. majus* var. JI7 allele of *ROS-EL* (**Figure 3.30**). As **Table 3.2** shows, the relative numbers of each *ROS-EL* allele in pool 2 was different from what was predicted from the original genotyping. Analyses that use this pool (A, D and E) are inaccurate at the *ROS-EL* locus. It is likely that, without eight JI7 *ROS-EL* alleles in pool 2, the peak at chromosome 6 in analyses D and E in **Figure 3.29** would be higher. This would also likely eliminate the small signal on chromosome 6 in analysis A.

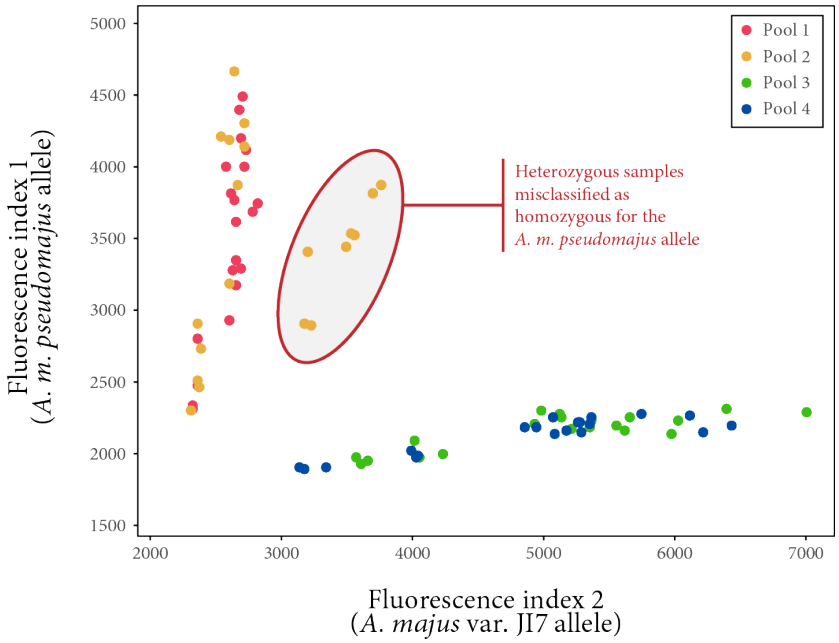


Figure 3.30 Scatter plot showing the *ROS-EL* genotyping results for the 80 L124 plants used for the second round of bulked segregant analysis in this chapter. The *y* axis shows the intensity of the green fluorophore seen when the *A. m. pseudomajus* allele has been amplified and the *x* axis shows the intensity of the blue fluorophore seen when the *A. majus* var. JI7 allele has been amplified. The points are coloured according to the pool to which they were assigned, as shown in the legend. The circled points are individuals that were misclassified as homozygous for the *A. m. pseudomajus* allele whereas they are in fact heterozygous.

Table 3.2 Intended and actual numbers of *A. m. pseudomajus* and JI7 alleles of *ROS-EL* in the four pools used in the bulked segregant analysis for L124.

	Intended number of alleles		Actual number of alleles	
Pool	*A. m. pseudomajus*	JI7	*A. m. pseudomajus*	JI7
1	40	0	40	0
2	40	0	32	8
3	0	40	0	40
4	0	40	0	40

Notably missing in **Figure 3.29** are the peaks on chromosomes 2, 5 and 8 seen in the previous analysis (**Figure 3.16**). The two differences between J108, used for the first BSA, and L124, used for the second, were the sizes of the families and the phenotyping method used. J108 was pooled according to differences in the white face phenotype as observed on living plants and from unprocessed photographs. L124 was pooled

according to differences in the white face phenotype as observed from the colour-converted images, largely discounting variation in yellow pigmentation. The first method resulted in several distinct peaks; the second method results in just one peak. This suggests that the peaks seen in the previous analysis were not linked to the white face phenotype, but instead were linked to additional variation between the pools.

My results from this experiment indicate that hypothesis 1 is correct – that only one gene is involved in changing the white face phenotype that segregates between these pools. However, this results also disproves the original hypothesis that a novel gene would be involved in changing this phenotype. It appears instead that the *ROSEA-ELUTA* locus is involved in regulating magenta pigmentation in a novel way.

3.6 Results: Genotyping a larger population maps the white face phenotype to the *ROSEA* locus

Bulked segregant analysis of L124 showed that the white face phenotype was likely genetically-regulated only at the *ROS-EL* locus. This meant that the additional phenotypic variation observed was due to genotyping errors or environmental variation. To determine which of these was responsible for the variation, I used the same two SNP markers at *ROS* and *EL*, as well as an additional marker at each gene (at 52,352,815 bp in the third exon of *ROS3* and at 52,491,840 bp, again in the *EL* exon), to genotype two additional F2 families grown in 2017: N101 and N102. These were generated using the same JI7 × *A. m. pseudomajus* accession cross as J108 and L124, but from a separate F1 individual (**Figure 3.31**). The white face phenotype was seen again in both families, with 24% (155 individuals) showing no white face (scores 0 and 1), 56% (363 individuals) having a weak white face (scores 2 and 3) and 19% (124 individuals) a strong white face (scores 4 and 5) (**Figure 3.32**). Once again, the ratios between phenotype groups appeared to resemble a 1:2:1 ratio, but a *G* test for goodness of fit gave $p < 0.01$, showing a significant difference between the expected and observed ratios.

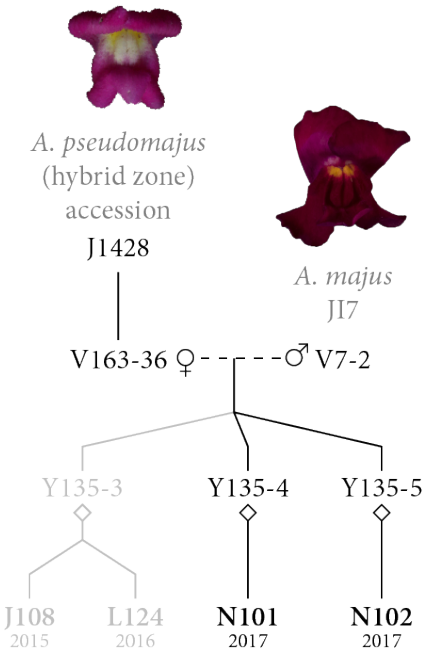


Figure 3.31 A pedigree of the two plant families grown in 2017 for analysis of the white face phenotype. Seed from J1428 was collected near the village of Ventola, Ribes de Freser, Girona, Catalonia and grown at the John Innes Centre as V163. Plant 36 from that family was crossed to the *A. majus* research line JI7 (individual V7-2). Plants 4 and 5 from the F1 generation (Y135) were each selfed to generate the families N101 and N102, respectively. These F1 plants were siblings to the one that gave rise to J108. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

Figure 3.32 Frequency of each white face phenotype score in the combined families N101 and N102 (n = 642), grouped according to phenotype description. The diagrams below the bars show a typical visual representation of each score.

As with L124, there was an association in N101 and N102 between the white face phenotype seen and the genotypes of *ROS* and *EL* as determined at the same marker sites as previously. Of the group with the strongest white face phenotype (score 5), all individuals had the genotype ROS^{P}/ROS^{P} and 80% had the genotype EL^{P}/EL^{P}; the *EL* genotype exceptions were heterozygous (EL^{P}/EL^{7}). Of those grouped as having the clearest non-white face phenotype (score 0), 58% had the genotype ROS^{7}/ROS^{7} and 60% had the genotype EL^{7}/EL^{7}; exceptions, again, were heterozygous (EL^{P}/EL^{7}). These results are shown in more detail in **Figure 3.33**. Both markers in both genes were consistent with each other.

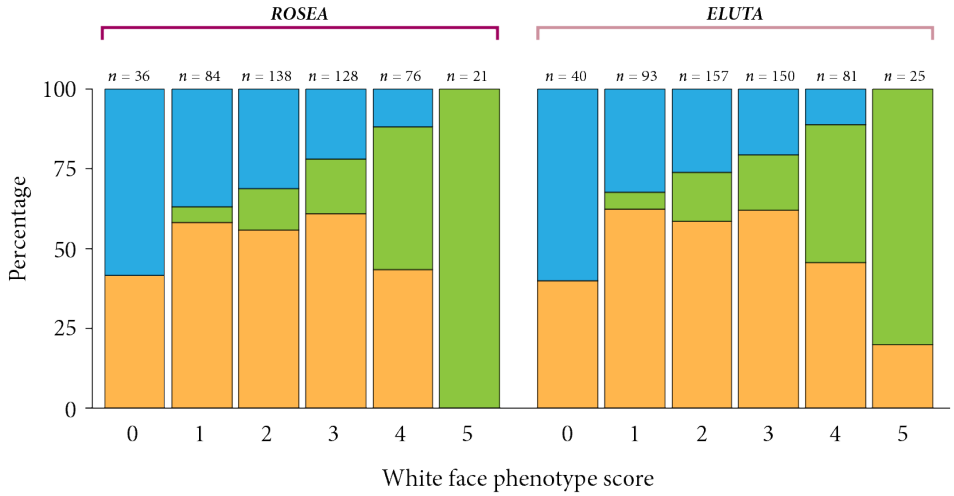


Figure 3.33 Percentages of each of the three *ROS* and *EL* genotypes present in individuals from the combined families N101 and N102 grouped by white face phenotype score. Individuals homozygous for the JI7 allele are coloured in blue, those homozygous for the *A. m. pseudomajus* allele are coloured in green and heterozygous individuals are shown in yellow. The number of individuals falling within each group is shown above its bar. These are not the same for both genes because genotypes could not confidently be called for all reactions.

In total, 25 recombination events were identified between *ROS* and *EL*, out of 424 plants whose genotypes were confidently called for all four markers. This translates to a genetic distance of 6 cM, which is more than 10 times higher than the previously described value of 0.5 cM using 10,261 individuals. This could be a sign that some of the individuals in the N101 and N102 populations were misgenotyped. The phenotypes of these recombinants are shown in **Figure 3.34**. Six of the seven individuals identified as homozygous for ROS^{P} but heterozygous at *EL* had a strong white face. The individuals identified as homozygous for ROS^{7} and heterozygous at *EL* had either weak white face or no white face phenotypes. Two of the three individuals homozygous for EL^{P} but heterozygous at *ROS* also had a strong white face, while the third had a weak white face. All those heterozygous at *ROS* and homozygous for EL^{7} had a weak white face. That there is a higher proportion of ROS^{P} homozygous recombinants (fifth row in **Figure 3.34**) showing a strong white face than EL^{P} homozygous recombinants (fourth row in **Figure 3.34**) suggest that the white face trait is linked more closely to *ROSEA* than to *ELUTA*. However, given that fewer EL^{P} homozygous recombinants were identified, and that the number of recombinants seen is higher than expected, this result is inconclusive.

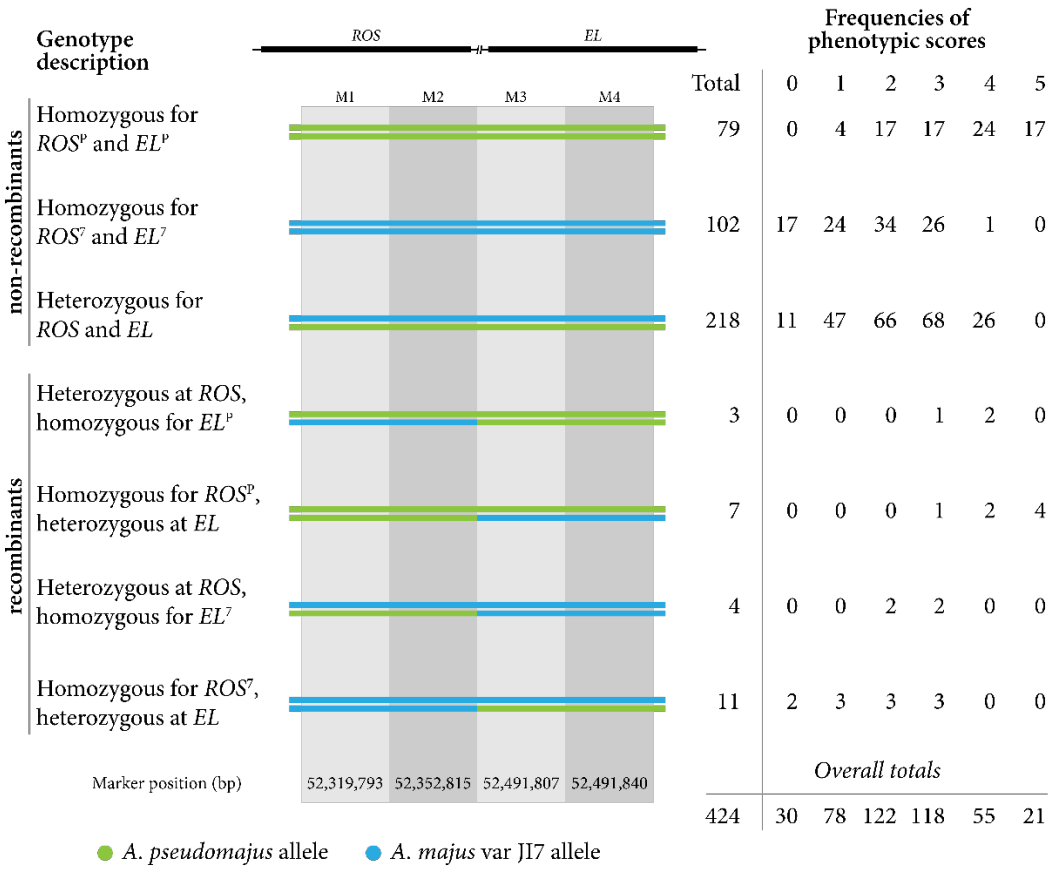


Figure 3.34 Phenotypic analysis of individuals with various genotypes at *ROS* and *EL*. Two markers are shown for each gene. The colour of the bars shows which allele or alleles were detected for an example individual at each of the four markers – green for *A. m. pseudomajus* and blue for JI7. The strongest white face phenotype is only seen when an individual is homozygous for the *A. m. pseudomajus* allele of *ROS*.

This likely overestimation in the number of recombinants may have occurred because of issues with the markers used. Markers I designed based on the genome sequencing data from J108 consistently performed poorer than those designed by LGC Ltd using their proprietary marker-design technology. Using the markers designed by LGC, nearly all reactions clustered into one of three classes, with fewer than 10 exceptions in each 96 well plate. The markers I designed, however, consistently resulted in 30 or more reactions not clustering with others and whose genotypes could not be called as a result (**Figure 3.35**). Reactions using LGC primers also clustered together more tightly than the ones using my manually-designed primers. This suggests that manually-designed primers are less efficient and less accurate than ones designed using the company's proprietary technology.

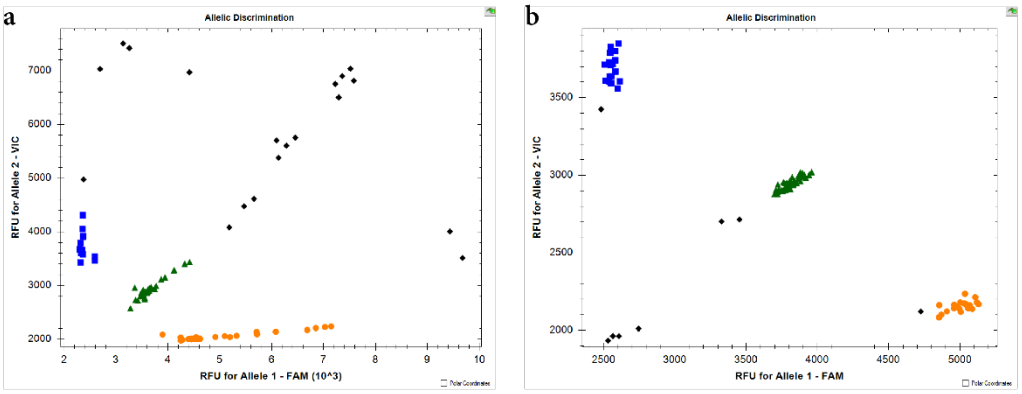


Figure 3.35 Allelic discrimination graphs used to determine the genotypes of a 96 well plate or reactions using a manually-designed marker for *EL* (**a**) and a marker for *FLA* designed by LGC Ltd using their proprietary technology (**b**). The *x* axis value for a reaction is the relative fluorescence intensity measured in the blue channel, detecting the FAM fluorophore (showing the presence of the JI7 allele), and the *y* axis value is the relative fluorescence intensity measured in the green channel, detecting the VIC fluorophore (showing the presence of the *A. m. pseudomajus* allele). Individuals determined to be homozygous for the JI7 allele are shown as orange circles, those determined to be homozygous for the *A. m. pseudomajus* allele are shown as blue squares and those determined to be heterozygous are shown as green triangles. Black diamonds represent reactions whose values did not cluster with any of the three classes and thus whose genotypes were not called. Plots taken from the CFX Manager 3.1 software (Bio-Rad Laboratories, Hercules, CA, US).

The ten putative recombinants homozygous for either ROS^P or EL^P were reserved at the end of the growing season and self-fertilised so that the phenotypes and genotypes of the resulting populations can be looked at in more detail in future experiments. If they are true recombinants, this would suggest that the recombination rates in JI7 or *A. m. pseudomajus* are much higher than previously estimated.

3.7 **Future experiments:** analysis of *ROS* and *EL* transcription in white face and non-white face flowers

One way of determining whether *ROS* or *EL* is responsible for the white face phenotype will be to look at which, if any, of the two is expressed in the white face patch and outside this region. If *ROS* controls the trait, its expression would be expected outside the white face patch, but not inside, given that *ROS* is an activator of magenta pigmentation. If *EL* is responsible for the trait, its expression would only be

expected inside the white face patch, and not outside, because *EL* is a repressor of magenta.

I selected plants from N101 and N102 classed as having a strong white face and no white face. I harvested the corollas of developing buds (0.5-1 cm in length) from the flowers of these plants and dissected them, collecting the ventral lip region (where the white face patch is seen) and the ventral and lateral lobes (which is magenta whether or not the white face phenotype is seen), pooling these together from several individuals (**Figure 3.36**). I extracted RNA from these, and this is currently being sequenced.

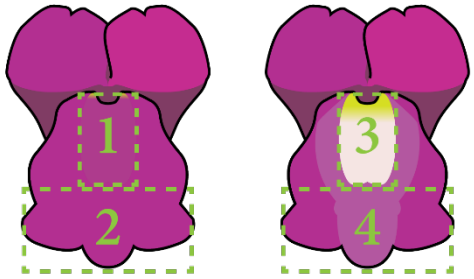


Figure 3.36 *Antirrhinum* flower diagrams showing the four tissue pools collected for RNA sequencing: (1) ventral lip region of a non-white face flower; (2) ventral and lateral lobes of a non-white face flower; (3) ventral lip region of a white face flower; (4) ventral and lateral lobes of a white face flower. The tissue was collected from developing buds but flowers are shown here in their fully developed forms for illustrative purposes.

Sequenced transcripts from this experiment will be mapped to the *Antirrhinum* genome and gene expression will be compared between the pools. These data will be used to test whether *ROS* or *EL* are expressed differently between the four tissues. One hypothesis for the white face phenotype is that *ROS* is downregulated in the ventral lip region of white face flowers. If this is the case, fewer sequenced transcripts would map to loci within the *ROS* genomic interval in the white face tissue pool (pool 3 in **Figure 3.36**) compared to the other three pools. Another hypothesis is that *EL* is upregulated in the ventral lip region of white face flowers, thus inhibiting production of anthocyanins here. If this hypothesis were correct, more sequence transcripts would map to loci within the *EL* genomic interval in the white face tissue pool (pool 3 in **Figure 3.36**) compared to the other three pools. Because *ROS* upregulates genes in the anthocyanin biosynthetic pathway and *EL* downregulates the same genes, I would

expect these biosynthetic genes to show lower levels of expression in the white face tissue whichever hypothesis is correct.

3.8 **Results:** Bulked segregant analysis of the white face trait from *A. molle*

Antirrhinum molle, which grows near the border between France and Catalonia, has white flowers with yellow foci and magenta veins. However, when this species was crossed to JI7, as is routinely done when new accessions are grown, a white face phenotype was produced. *A. molle* has a habitat range that neighbours (and, in part of its range, overlaps with) that of *A. m. pseudomajus*, and putative hybrids have been observed. This suggests that *A. molle* – or at least the accession used in this cross, which was collected where both species are found in sympatry – carries a white face allele from *A. m. pseudomajus*. Alternatively, the white face of *A. m. pseudomajus* may be the result of introgressive hybridisation with *A. molle*.

In some plants in the F2 between *A. molle* and JI7, there was a clear white face at the centre of the flower (**Figure 3.37 b**) whereas in others, this white face covered most of the upper half of the ventral and lateral lobes (**Figure 3.37 c**). I will refer to this as a 'white band'. Others had no white face phenotype (**Figure 3.37 a**), while a fourth group had the *rosea*^{dorsea} (*ros*^{dor}) phenotype (**Figure 3.37 d**), where magenta pigmentation is missing from the flower face although some pale pink colour is seen on the upper surface of the dorsal lobes.

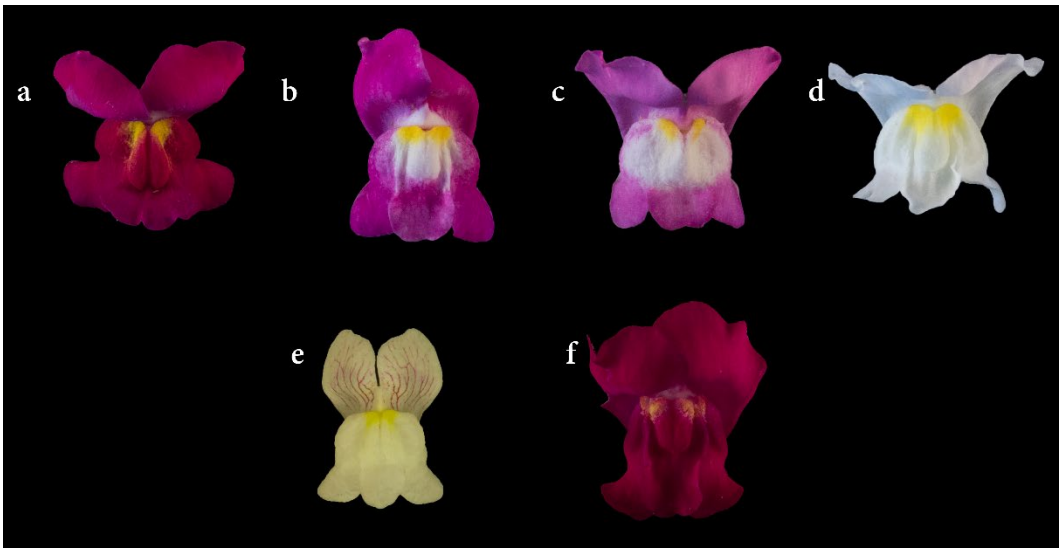


Figure 3.37 Phenotypes seen in J104 (**a** to **d**), an F2 from a cross between *A. molle* (**e**) and JI7 (**f**). In some plants, flowers had full magenta pigmentation with no white face (**a**), in some, they had a strong white face (**b**) and in some, flowers were white across the upper half of the ventral and lateral lobes ('white band', **c**). A further group of plants had a similar phenotype to that seen in *rosea*^{dorsea} mutants – barely any magenta pigmentation (**d**).

An *A. molle* individual was crossed to JI7 and an F2 population of 500 plants (J104) was produced by selfing the F1 (**Figure 3.38**). The *A. molle* individual came from the C-QUE accession (**Figure 3.39**). Of the 465 F2 plants whose flower colours were scored, 230 (49.4%) had full magenta pigmentation (no white face), 126 (27.1%) had a strong white face, 80 (17.2%) had a *ros*^{dor} phenotype and 29 (6.2%) had a white band phenotype. The number of white band individuals is one 16th of the total number of individuals scored, which may mean that it is homozygous for recessive alleles at two independent loci that regulate the phenotype. I tested whether the ratios seen adhered to the 9:3:3:1 phenotypic ratio that would be expected for two unlinked segregating loci with a dominant allele at each. However, a *G* test of goodness of fit showed that 230:126:80:29 differed significantly from that ratio ($p < 0.001$). I also looked whether the magenta (regardless of white face phenotype) to *ros*^{dor} phenotype ratio was 3:1 as it is for plants segregating for the *ros*^{dor} allele in JI7. However, the observed ratio again differs significantly from the expected ratio ($p < 0.001$). Finally, I tested whether the ratio of the combined no white face and centralised white face plants to the combined *ros*^{dor} and white band plants adhered to a 3:1 ratio. A *G* test of goodness of fit showed that the observed 256:109 ratio did not differ significantly from the expected ratio ($p = 0.434$). Moreover, the ratio of *ros*^{dor} to white band plants is also not significantly different from a 3:1 ratio ($p = 0.701$). This led to a hypothesis that two genes regulate

magenta in *A. molle*, although some phenotypic plasticity may make scoring intermediate phenotypes (possibly heterozygotes) difficult. According to this hypothesis, plants with a white band are homozygous for the recessive *A. molle* allele at both genes. An alternative hypothesis is that only one gene regulates magenta variation and that differences in the shape and size of the white pattern are all because of environmental effects.

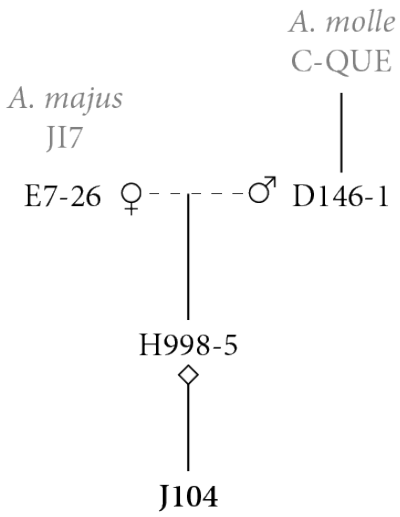


Figure 3.38 A pedigree of J104, a plant family grown for analysis of the flower colour variation from *A. molle*, which included a white face phenotype. This family originated from a cross between the plant D146-1, which was generated from *A. molle* seed collected from the C-QUE location, and an individual from the JI7 lab cultivar (E7-26). In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

Figure 3.39 Location within Catalonia of the C-QUE collection location where *A. molle* was sampled in 2003. The location was named after the Santuari de Santa Maria de Queralt near Berga, close to where the accession was sampled.

To test these two hypotheses, the plants were pooled according to phenotype for bulked segregant analysis. Three bulks were constructed: the first bulk, full magenta, contained plants without any white face pattern seen on the flowers; the second bulk, white face, contained plants with a white face pattern that did not extend beyond the central region of the flower; and the third bulk, white band, contained plants whose flowers had a white patch that extended into the lateral lobes of the flower. DNA from plants in each bulk was pooled and sequenced by Yongbiao Xue at the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. I processed the results using the same pipeline I had used for J108 and L124 (white face from *A. m. pseudomajus*) and looked for peaks of allele frequency differences by calculating *G′* values and Δ SNP-index values for windowed SNPs across the genome. I performed these analyses for three phenotypic comparisons: full magenta v centralised white face; full magenta v spread white band; and centralised white face v spread white band (**Table 3.3**).

Table 3.3 Predicted results from bulked segregant analysis for three comparisons in J104 given two possible hypotheses: that variation in white face size and shape is genetic and that two loci regulate the phenotype; or that variation in white face size and shape is mostly environmental and that magenta colour in this family is only regulated at one locus.

Comparison	Prediction if hypothesis 1 is correct	Prediction if hypothesis 2 is correct
Full magenta v centralised white face	One peak	One peak
Full magenta v spread white band	Two peaks	One peak (the same peak as in the first comparison)
Centralised white face v spread white band	One peak (but not the same peak as in the first comparison)	One peak (the same peak as in the other comparisons)

The BSA comparing plants that had full magenta flowers with those that had strong but centralised white face patterns on their flowers showed just one peak in G' value, which was located on chromosome 2 (**Figure 3.40**). The narrow peaks on chromosomes 3, 4 and 5 were not considered for the same reasons given for J108 earlier in this chapter. The peak on chromosome 2 suggests that the white face pattern seen in this family is regulated by a gene on this chromosome and does not have the same regulatory basis as the white face phenotype seen in *A. m. pseudomajus*. *CHI*, an anthocyanin structural gene, is encoded on chromosome 2, although from this data it is not possible to determine whether this or another locus on the ~70Mb-long interval is causal to the phenotype.

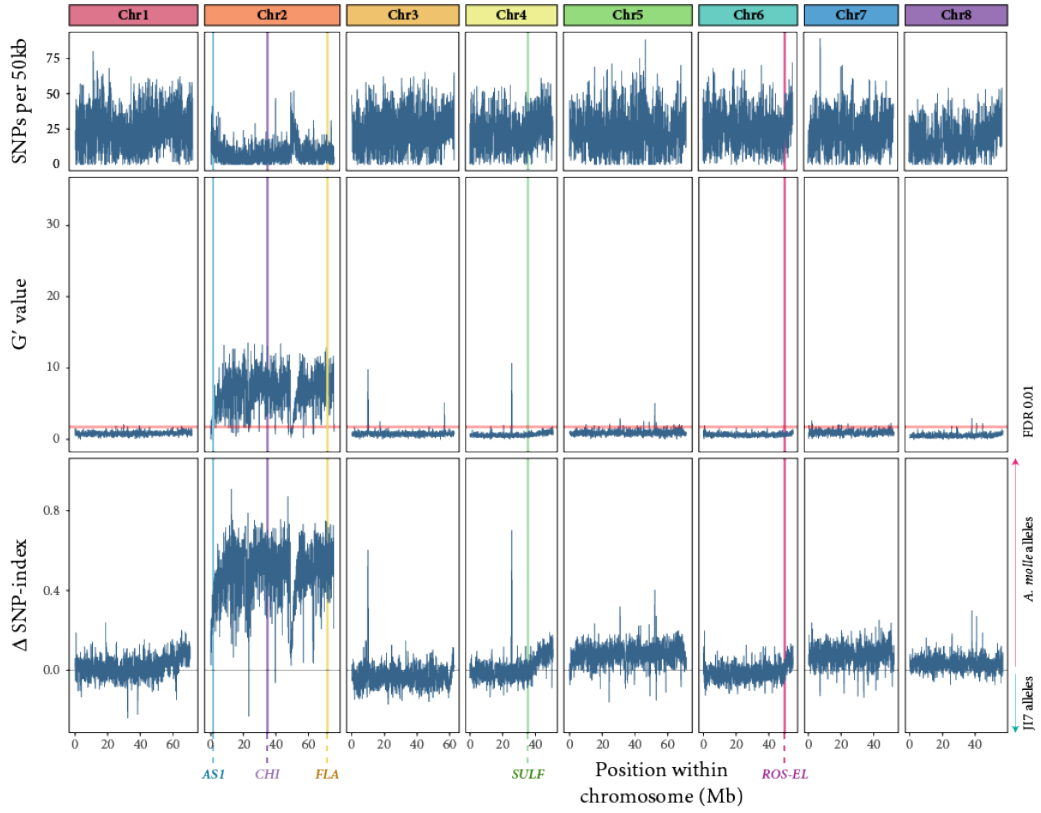


Figure 3.40 Bulked segregant analysis Manhattan plots for family J104, comparing allele frequencies in plants that have full magenta with those that have a strong centralised white face. The top plot shows the number of SNPs in each window, the middle plot shows the *G′* value and the bottom plot shows the difference in allele frequency between the pools. All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a *G′* threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *CHI*, *FLA*, *SULF* and *ROS-EL* are indicated with vertical lines and labelled below the *x* axis.

The second BSA of this family, comparing plants that had full magenta flowers with those that had a spread white band pattern showed two peaks (**Figure 3.41**). The first was the same chromosome 2 peak seen in the previous comparison. However, the *G′* values for this peak were much higher in the second comparison than in the first (maximum values of around 32 and 12, respectively). This suggested that alleles linked to the white face/band phenotype were fixed in the white band pool but not in the centralised white face pool – ie plants with a centralised white face are heterozygous, while those that have a white band are homozygous for the *A. molle* allele.

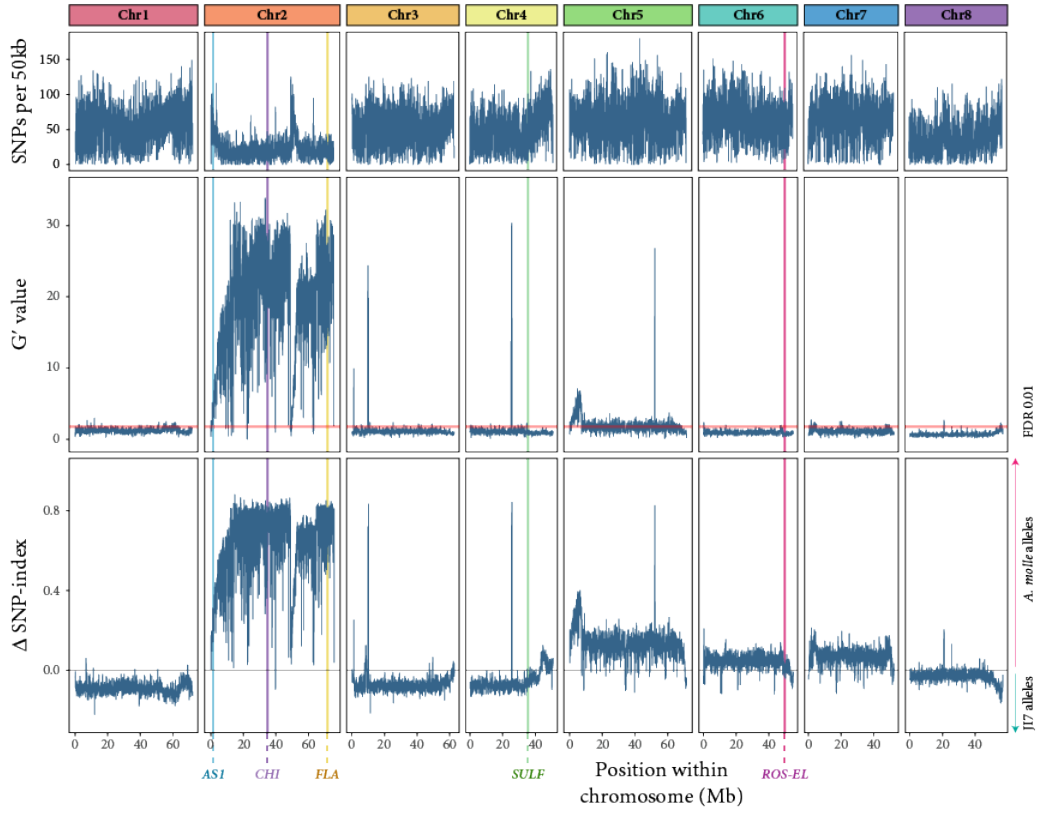


Figure 3.41 Bulked segregant analysis Manhattan plots for family J104, comparing allele frequencies in plants that have full magenta with those that have a spread white band pattern. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools. All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA*, *SULF* and *ROS-EL* are indicated with vertical lines and labelled below the *x* axis.

The second peak seen was near the start of chromosome 5, between the ~3Mb and ~7Mb positions. The presence of this peak in this comparison (no white v white band covering the upper half of the lower lobes) and not in the previous one (no white v centralised white face) suggested that a modifier of the white face phenotype may be encoded in this interval. However, this peak was low compared to that seen on chromosome 2, suggesting that the *A. molle* allele was only marginally more prominent in the white face pool compared to the non-white face pool. This may mean that the pools varied in a trait regulated by a gene encoded at this interval because of the small sizes of the bulks. I also investigated whether this was the same peak seen in J108 comparing white face and non-white face pools in *A. m. pseudomajus* × JI7 F2, but the

J104 peak appears to be closer towards the 3′ end of chromosome than the J108 peak (**Figure 3.42**).

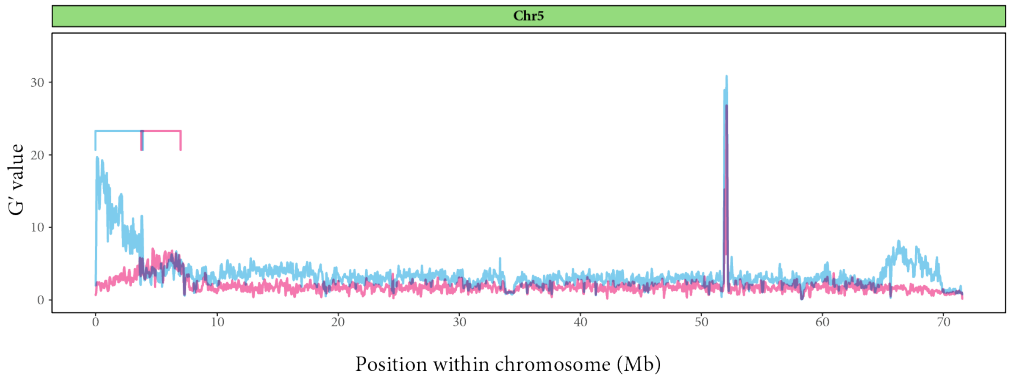


Figure 3.42 A closeup view of chromosome 5, showing the *G*′ value profiles for J108 comparing the white face and full magenta phenotypes (blue line) and J104 comparing the strong white band and full magenta phenotypes (pink line). J108 was the F2 generation from a cross between *A. m. pseudomajus* and JI7. J104 was the F2 generation from a cross between *A. molle* and JI7. The coloured brackets above the line plot show the width of the peaks in each comparison in their respective colours.

As a final comparison of this family, I looked at the allele frequency differences between the centralised white face pool and the white band pool. This showed the same peaks as in the comparison of the full magenta (non-white face) pool and the white band pool, but at lower *G*′ values (**Figure 3.43**).

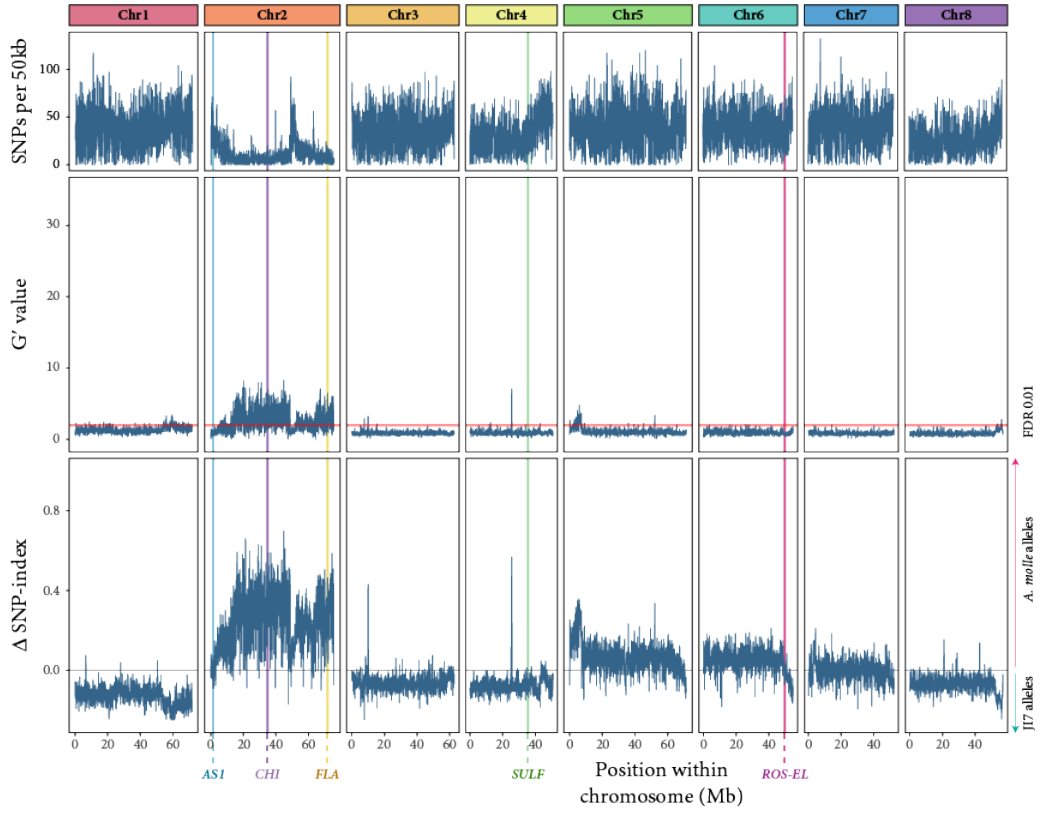


Figure 3.43 Bulked segregant analysis Manhattan plots for family J104, comparing allele frequencies in plants that have a centralised white face pattern on their flowers with those that have a spread white band pattern. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools. All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA*, *SULF* and *ROS-EL* are indicated with vertical lines and labelled below the *x* axis.

I looked at the sequencing coverage seen for these samples at the *ROS* locus and compared this to the coverage for individuals of *A. m. pseudomajus*, *A. m. striatum* and *A. molle* (**Figure 3.44**). *A. molle* individuals have a deletion in the promoter region of *ROS1*, a deletion also seen in *A. m. striatum*. However, the deletion is absent from the sequenced samples in J104, suggesting that it was also absent in the accession used from C-QUE, which has not been sequenced. This may mean that the *A. molle* accession used for this analysis had the *ROS* allele of *A. m. pseudomajus*, the result of hybridisation at the C-QUE location.

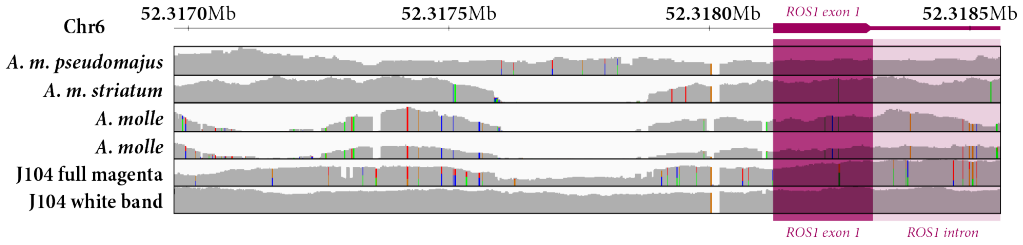


Figure 3.44 Whole genome sequencing depth of coverage along a section of chromosome 6 containing the first exon of *ROS1* from *A. m. pseudomajus*, *A. m. striatum*, *A. molle* and two pools from J104. The sample names are shown on the left, and their coverage profiles are illustrated using panels from the Integrative Genomics Viewer (Robinson *et al* 2011). The location of the *ROS1* exon is shown in magenta, and the intron that follows it is shown in light pink.

Because a similar peak on chromosome 2 was first seen in the F2 from the cross between *A. m. pseudomajus* and JI7, and that this was likely a result of variation in yellow pigmentation, I looked back at the scoring data for J104 flowers to see whether there was variation in yellow pigmentation. *A. molle* yellow colour restriction is known to be regulated by *FLA* rather than *SULF* (Boell *et al* unpublished results), and therefore yellow variation in J104 would be expected to map to chromosome 2. For yellow colour phenotyping, J104 flowers had been scored as having 'very restricted yellow' (189 plants), 'slightly restricted yellow' (158 plants) or 'strong yellow' (111 plants) flower face/foci phenotypes. This is close to a 3:1 ratio of yellow restriction to strong yellow, and a *G* test for goodness of fit showed that there was no significant difference between the observed ratio and 3:1 (p = 0.786). This suggested that one locus controlling yellow restriction was segregating in the population and that the *A. molle* allele is dominant to that of JI7.

Looking at the individuals used in the pools, 88% of those in the 'white band' pool and 72% of those in the 'centralised white face' pool had very restricted yellow, compared with just 24% in the 'full magenta' pool. This suggests that there was a strong difference in the number of *A. molle FLA* alleles in the pools. The reason for this difference in yellow could be because of two things: either there was unconscious bias in selecting the plants that made up these pools, as in J108; or, given that there was no other peak in the BSA of J104, the white face regulator in this family may be genetically linked to *FLA*, and individuals with a strong white face were more likely to have restricted yellow. I looked at the ratios of yellow phenotypes within each magenta class (**Table 3.4**) and compared these to the overall observed phenotypic

ratio. In each class, *G* tests for goodness of fit showed that the within-magenta class ratio of yellow phenotypes differed significantly from the overall ratio of yellow phenotypes ($p < 0.01$ for all four magenta classes). This suggests that there is some linkage between the yellow and magenta phenotypes seen in J104. Alternatively, the two different segregating phenotypes made scoring each one independently difficult. As there are no photos from most J104 individuals, testing this will require another segregating population.

Table 3.4 Number of individuals assigned to each magenta and yellow phenotypic class in J108. These number differ slightly from those given for magenta phenotype at the start of this section because some individuals were not scored for yellow.

	Very restricted	Slightly restricted	Strong yellow	**Total**
Full magenta	54	70	104	**228**
White face	74	45	7	**126**
White band	17	8	0	**25**
ros^{dor}	44	35	0	**79**
Total	**189**	**158**	**111**	

3.9 Discussion

3.9.1 The *ROS-EL* locus produces at least three phenotypes in the *Antirrhinum* genus

I started this chapter with a hypothesis that the white face phenotype was caused by an *A. m. pseudomajus* allele at a single gene, similarly to the other magenta phenotypes studied in *Antirrhinum*. My results suggest that this hypothesis is correct, and that only one locus underlies the phenotype. However, contrary to my initial expectations, this locus is not a previously unidentified gene, but appears to be either *ROSEA* or *ELUTA*, both of which are previously known regulators of magenta pigmentation.

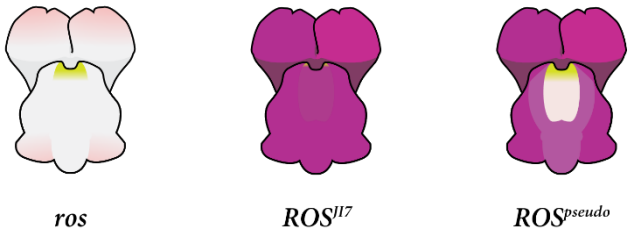


Figure 3.45 Schematic diagrams of the phenotypic effects of three different alleles of *ROSEA* in a common JI7 background according to the results presented in this chapter. When the non-functional *ros* allele (from a JI7-related cultivar) is homozygous, very little anthocyanin is produced in the petals and, as a result, they appear a very pale pink. When the *ROS* allele of *A. majus* var. JI7 (ROS^{JI7}) is homozygous, anthocyanin is produced throughout the petals, giving the flowers a bright magenta hue. When the *A. m. pseudomajus* allele of *ROS* (ROS^{pseudo}) is homozygous, the same bright pigmentation is produced as with ROS^{JI7}; here, however, a patch on the flower face is left without pigmentation, leading to the white face phenotype seen in *A. m. pseudomajus*.

Given these results, one of these genes appears to be responsible for regulating anthocyanin production in *Antirrhinum* in a way not previously described. Previous work has shown that plants carrying the functional *ROS* allele accumulate anthocyanin pigment throughout their petals, whereas *ros* mutants show very little anthocyanin accumulation (Schwinn *et al* 2006). But from my results, if the white face phenotype in my segregating families is regulated by the *A. m. pseudomajus* allele of *ROS*, this allele would seem to extend magenta pigmentation throughout most of the flower, but leaving part of the flower without accumulating anthocyanin pigment, thus producing the white face phenotype. Likewise, the *A. m. pseudomajus* allele of *EL* was thought to be non-functional, as *A. m. pseudomajus* has magenta pigmentation that is not restricted to the centre of the flower, as is seen with the *A. m. striatum* allele. My results suggest that, if EL^{P} is responsible for the white face phenotype, this allele is functional but is regulating magenta in a different part of the flower from the *A. m. striatum* allele.

One potential explanation for *ROS* or *EL* having these two distinct effects could be presence of enhancers located in the *cis*-regulatory region of the gene. In *Heliconius* butterflies, wing colour patterns are regulated by transcription factors and, similarly to petal colour regulation in *Antirrhinum*, only a small number of genes of large effect are involved in this regulation. Red wing colour patterns in *Heliconius* are regulated by the transcription factor *OPTIX*, and variation in the red pattern consistently maps to this

locus (Reed *et al* 2011). Two contrasting red wing pattern phenotypes seen in *Heliconius* are *DENNIS* and *RAY*, which have red bands on the top and bottom halves of the wings, respectively. Using genomic analyses, Wallbank *et al* (2016) showed that these phenotypes are associated with a two-part 50 kb-long sequence located 60-110 kb 3′ of the *OPTIX* gene. The first part of this sequence is associated with *DENNIS* and the latter part with *RAY* (**Figure 3.46**). A similar *cis*-regulatory mechanism of producing magenta patterns by regulating *ROS* or *EL* transcription may be found in *Antirrhinum*. In such a system, some species, such as *A. m. pseudomajus*, would carry a *cis*-regulatory sequence near *ROS* or *EL* which differs from the sequence at the same position in other species. Future work could make use of natural populations of species with full magenta but no strong white face to look for phenotype-genotype associations.

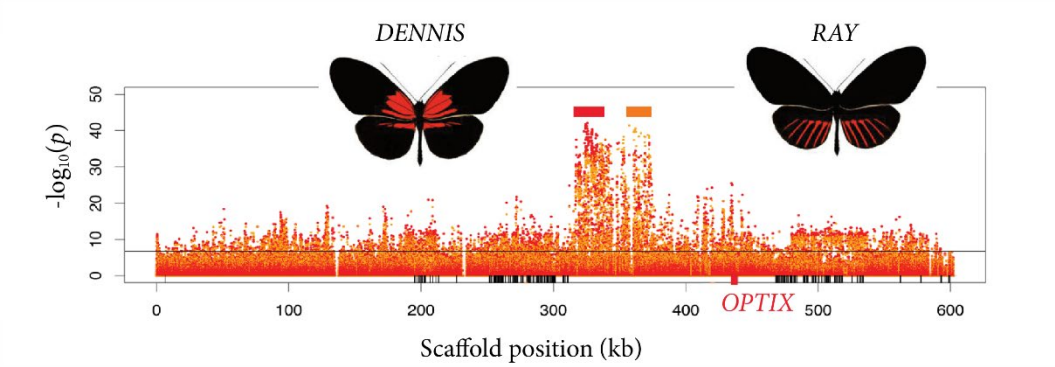


Figure 3.46 Association study of wing patterning in *Heliconius* butterflies. The inset butterfly diagrams show the *DENNIS* and *RAY* wing pattern phenotypes. The plot shows the statistical association across 96 genomes for the *DENNIS* phenotype (red dots) and the *RAY* phenotype (orange dots). These regulatory sequences, whose locations along the chromosome are shown with rectangles above the points (red for *DENNIS*, orange for *RAY*), interact with the *OPTIX* gene, whose location is shown under the plot using a red rectangle. Adapted from Wallbank *et al* (2016).

3.9.2 The white face phenotype seen in crosses between JI7 and *A. molle* is not the same trait as the white face from *A. m. pseudomajus*

When an *A. molle* individual was crossed to JI7, the F2 population segregated for a trait similar to those seen in F2s from crosses between JI7 and *A. m. pseudomajus*. This suggested that *A. molle* (or at least the C-QUE accession used in this cross) also carried the white face allele at *ROS-EL*, perhaps through hybridisation with *A. m. pseudomajus*, but that its phenotype was not seen because *A. molle* has no anthocyanin pigmentation outside the dorsal veins region. However, when white face and non-white face

individuals were compared using BSA, there was no difference in allele frequency on chromosome 6, where the *ROS-EL* locus is located. The variation between the pools instead mapped to chromosome 2, which had not previously been implicated in interspecies variation in magenta colour, suggesting that *A. molle* at C-QUE have a magenta-regulating mechanism not yet described.

One known flower colour gene on chromosome 2 is *CHALCONE ISOMERASE* (*CHI*), which encodes the first enzyme involved in the conversion of chalcone to anthocyanins. Because *A. molle* has a magenta venation pattern in its flowers, and because flavonoids produced using *CHI* are important for plant defence (Rausher 2006), *CHI* is unlikely to be non-functional in *A. molle*. But the *A. molle* allele of *CHI* may contain *cis*-regulatory changes that could affect the way it is expressed and in which tissues (Rausher 2006, Streisfeld and Rausher 2011). Alternatively, there may be a novel transcription factor-encoding gene located on chromosome 2 that regulates magenta colouration in the face of *Antirrhinum* flowers. It is not possible to address this further using the data generated using J104 because of the low recombination between *A. molle* (and several other species) and JI7 on that chromosome. One possible future experiment could be to cross *A. molle* from C-QUE to a magenta-flowered *Antirrhinum* accession without the white face phenotype and generating a segregating population. If the white face trait segregates in this population, it may be possible to map the phenotype more precisely.

3.9.3 Bulked segregant analysis can be problematic when two or more interacting traits segregate

Pooling individuals into bulks is an effective and efficient way of finding the genetic basis of phenotypic differences. The number of DNA preparations and sequencing lanes required is equal to the number of bulks analysed, making this a time- and cost-effective way of mapping traits. BSA in its simplest form also makes phenotyping a less onerous task as only individuals with the most extreme phenotypes need to be scored. It is an especially useful method for mapping traits determined by few genes of large effect, as flower colour typically is.

However, methods that rely on sequencing pooled DNA have their caveats. Once tissue or DNA from individuals is pooled, the ability to identify individuals is lost. DNA amplification during the library preparation stage of sequencing may not be even across individuals. Because of this, individuals may not be evenly represented in the

read counts at each locus, and allele frequency estimates may be incorrect. This problem can be minimised by optimising library preparation and using the highest expected coverage possible (Anderson *et al* 2014). A related technique called multiplexed shotgun genotyping (MSG) is now emerging as a successor to BSA. MSG has the advantage that individuals are barcoded during the library preparation stage before being pooled, giving the ability to identify individuals after sequencing (Andolfatto *et al* 2011).

Pooling individuals can also present problems because of variation in more than one trait. Some of this variation is likely to be in traits that are not visible or easily scored. The effect of this variation can be minimised by using pools containing as many individuals as is feasible (Magwene *et al* 2011). However, I encountered an additional source of such variation.

Although I did not notice at the time, the plant pools from family J108 used for the first round of bulked segregant analysis in this chapter differed not only for their magenta phenotype, but also for their yellow phenotype. Because of this additional variation, I had a peak in G' value on chromosome 2, a part of the genome that is not linked to the white face trait. This peak disappeared when plants in a related family were pooled based on their magenta phenotypes alone, by scoring from colour-adjusted photographs. This highlights the importance of eliminating unconscious bias when pooling individuals according to their flower colour phenotypes. Selecting plants with the most extreme phenotypes for one flower colour trait without accounting for other flower colour variation may introduce such bias.

I also selected the individuals for the first round of bulked segregant analysis by physically grouping plants that appeared to have the same phenotypes together. Because plants in this F2 population also showed variation for traits other than flower colour, this may have had the unintended consequence of grouping together plants with other traits in common, such as height, leaf shape or growth habit. This additional source of unconscious bias may explain the peaks on chromosomes 5 and 8 as these do not appear to be flower colour-related. These peaks are eliminated when pooling is done using only flower photographs, without being able to see the full plant.

3.9.4 Using small bulks can lead to false signals

In **Figure 3.16**, there is a residual difference in allele frequency seen across much of the genome outside the peaks on chromosomes 2, 5, 6 and 8. This implies some

phenotype-associated loci on these chromosomes in some of the individuals constituting the bulks. This could be another case of inadvertent selection, similar to the phenotyping bias that led to the errant peaks discussed previously. However, given that these differences in allele frequency are low, any phenotypic differences underlying them are unlikely to be fixed in more than 10% of the plants sampled. The bulks in these analyses only contained 20 individuals each. It is therefore possible that phenotypic differences in one or two plants, combined with linkage disequilibrium between *A. m. pseudomajus* and *A. majus* var. JI7, is responsible for these allele frequency differences. This could be tested by using a larger number of individuals in each bulk, where the effect of outlying individuals would be minimised.

3.9.5 Phenotypic plasticity can make mapping traits difficult

One of the difficulties I experienced with the white face trait was the variability of the phenotype. Although a strong association exists between the clearest white face phenotype and the *A. m. pseudomajus* allele of *ROS-EL*, it is often impossible to predict the genotype of plants with a weak white face phenotype. Because the plants were grown outside, they were subjected to a range of environmental conditions. Different individuals may have experienced different levels and sources of environmental stress, for example because of their position within the tray in which they were grown, their proximity to shade or the uniformity of watering. Anthocyanin production is a common stress response in plants (Christie *et al* 1994, Dixon and Paiva 1995, Miki *et al* 2015). The effect of environmental stresses on flower colour in *Antirrhinum* is not known, but stress may cause a plant-wide increase in anthocyanin accumulation, thus making flowers darker. Alternatively, a requirement to produce anthocyanins in vegetative tissue may limit production in flowers, leading to paler flowers.

3.9.6 Misgenotyping

One of the problems I encountered when carrying out this work was misgenotyping samples using KASP. In the graph in **Figure 3.30**, eight of the 20 non-white face plants I initially assigned a ROS^{P}/ROS^{P} genotype turned out to be ROS^{P}/ROS^{7} heterozygotes. This incorrect assignment of genotype occurred because the blue fluorescence of these samples, corresponding to the *A. majus* var. JI7 allele, was not as bright as for some of the other samples. As a result, these samples were incorrectly clustered with the ROS^{P}/ROS^{P} homozygotes. This result shows the importance of manually checking the

KASP output for each plate. For my subsequent results in section 3.6, I visually checked each genotype output graph to look for discrepancies between assigned genotype and actual genotype.

I also had to discount several samples genotyped using KASP. Samples located around the top edge of the 96-well PCR plates I used for these analyses often evaporated. This improved slightly by sealing the plates using a higher-quality foil lid, but some evaporation persisted. These empty wells were sometimes assigned a heterozygous genotype by the clustering algorithm. Most of these were easy to check by looking at the output graph manually, but I could not determine whether some samples were true heterozygotes or evaporated reactions. Because of this uncertainty, I eliminated all samples from the top row of each plate and reanalysed them separately.

It is also possible that some misgenotyped samples escaped after these manual adjustments. This could explain the likely exaggerated recombination rate calculated between *ROS* and *EL*. Although I used two markers for each gene, the positions of these putative recombinants within each genotyping plate may mean consistent misgenotyping of some samples. This could be avoided by collecting each sample twice and using a different layout in each replicate plate.

3.9.7 Implications for a previously studied *Antirrhinum* hybrid zone

ROS and *EL* have been studied extensively in a hybrid zone between *A. m. pseudomajus* and *A. m. striatum* with magenta and yellow flowers, respectively. In *A. m. pseudomajus*, *ROS* extends magenta pigmentation throughout the petals apart from the white face region; this subspecies is homozygous for recessive *el*. In *A. m. striatum*, *EL* restricts the magenta venation pattern regulated by the unlinked *VE* gene to the centre of the dorsal petals; *A. m. striatum* is homozygous for recessive *ros*. In the 1-2 km long hybrid zone between them, hybrid flowers with intermediate colours and patterns are found, but these phenotypes are confined to this geographic region (Whibley *et al* 2006, Tavares 2014).

A cline in *ROS-EL* allele frequency is found in this hybrid zone, with the *A. m. pseudomajus* haplotype increasing sharply in frequency from 0 to 1 over the length of the hybrid zone (**Figure 3.47**). A peak in F_{ST}, a measure of population differentiation, is also found at the *ROS-EL* locus. These are signs that the locus is under intense selective pressure because the contrasting phenotypic effects of *ros EL* and *ROS el* are crucial for the fitness of *A. m. striatum* and *A. m. pseudomajus*, respectively.

The white face phenotype is not expected to contribute to these signals of selection. The phenotype is likely of great importance to *A. m. pseudomajus* because it allows the yellow and magenta pigments of its flowers to separate to ensure clearer pollinator guides. But a reciprocal selective pressure may not occur in *A. m. striatum* because the regulator of the white face phenotype only has an effect if the flower is magenta. No effect would be expected on the yellow colour in *A. m. striatum*, although if *ELUTA* is the white face regulator, it may affect the magenta veins.

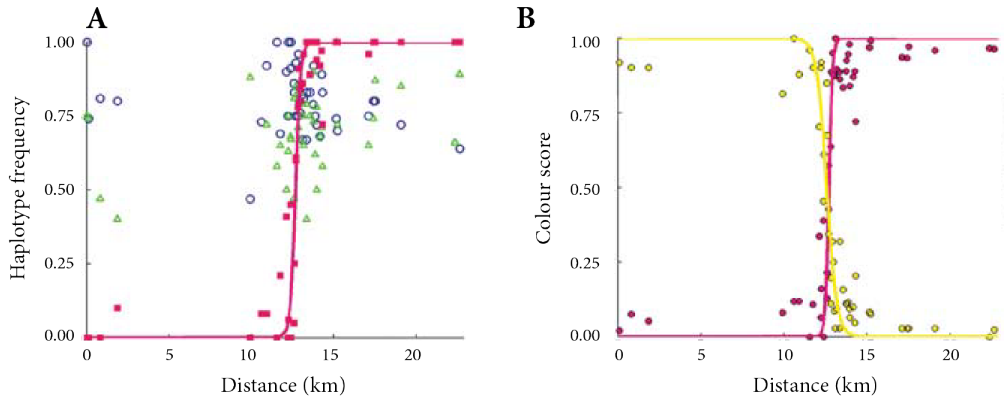


Figure 3.47 Clines in haplotype frequency and flower colour in a hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. Panel A shows the steep cline in the frequency of the *ROS1 A. m. pseudomajus* haplotype (magenta points and line) along the hybrid zone from east to west. Unlinked genes (green and blue points) do not show such a cline. Panel B shows the corresponding clines in yellow and magenta flower colour. Taken from Whibley *et al* (2006).

4 The aurone biosynthetic gene *FLAVIA* regulates yellow colour variation in *Antirrhinum majus*

4.1 Introduction

4.1.1 Yellow flower colour

Yellow pigmentation is common in flowers, especially those pollinated by bees. Bee-pollinated flowers often have yellow pigmentation throughout the flower, sometimes combined with nectar guides that use other pigments or structural colours to facilitate pollinator foraging (Wilson *et al* 2004). Yellow pigments themselves can also form these guides, and blue- and magenta-coloured flowers often contain patches of yellow pigmentation (Owen and Bradshaw 2011).

Several classes of compounds are responsible for yellow flower colours in plants. Bright yellow colours in most flowers are produced through the accumulation of highly oxidised carotenoids known as xanthophylls (Nakayama 2002, Glover 2014). Daffodils (*Narcissus* spp), oilseed rape (*Brassica napus*) and monkeyflowers (*Mimulus* spp) are examples of plants that use these carotenoids to make their flowers bright yellow (Valadon and Mummery 1968, Yuan *et al* 2014, Zhang *et al* 2015). Several

flavonoid compounds can also reflect light in the yellow part of the visible spectrum. Pale yellows in carnations (*Dianthus* spp) and cyclamens (*Cyclamen* spp) are produced by chalcones, while several members of the family Asteraceae use 6′-deoxy chalcones to produce their pale yellows. Brighter yellows require either supplementary carotenoid accumulation or the conversion of chalcones to aurones. Aurones are rarer than most other flavonoid pigments and are only found in a small number of species of flowering plants, including *Cosmos* (Asteraceae), *Limonium* (Plumbaginaceae) and *Antirrhinum* (Rausher 2006, Tanaka *et al* 2008). That these species are unrelated suggests convergent evolution of aurone synthesis, and it is not known if other species use the same biosynthetic pathway as *Antirrhinum*, where aurones have been best studied (Tanaka *et al* 2008).

4.1.2 Yellow flower colour in *Antirrhinum*

4.1.2.1 Aurone biosynthetic pathway

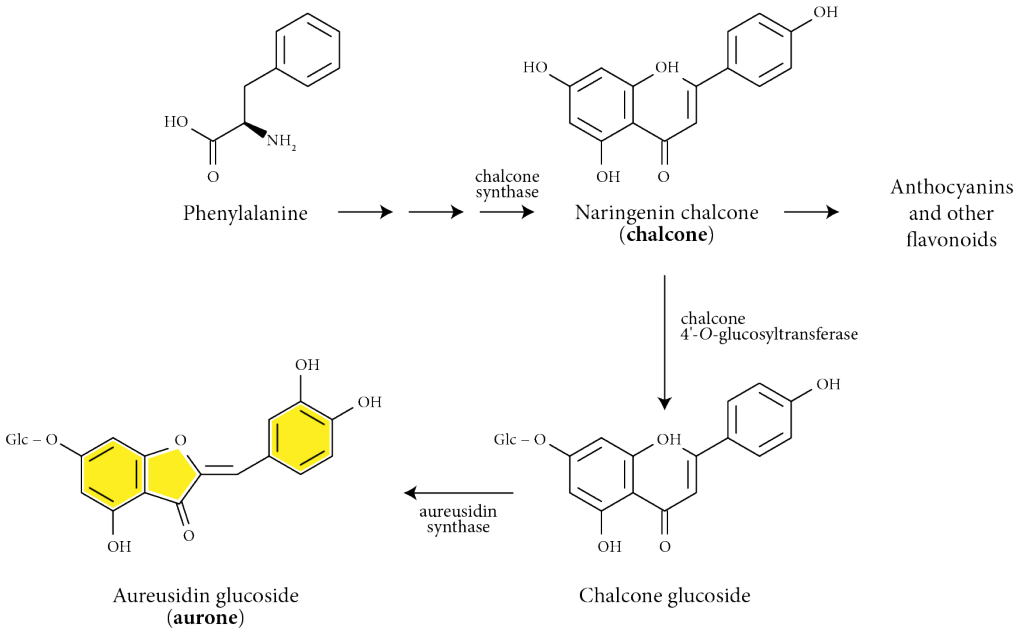


Figure 4.1 Simplified biosynthetic pathway of aureusidin glucoside, the yellow aurone pigment in *Antirrhinum* flowers. Like anthocyanins, aurones are derived from chalcone, but unlike the large pathway used to produce anthocyanins, the aurone biosynthetic pathway is relatively short. Just two enzymes are involved in converting chalcone to aurone, whereas anthocyanins require at least five. Adapted from Ono *et al* (2006).

Yellow aurones share a molecular precursor with anthocyanin in chalcone. Chalcone

(specifically 2′,4,4′,6′-tetrahydroxy chalcone, naringenin chalcone or THC) itself has a pale-yellow colour, but it is unstable and usually requires modification to be used as a pigment (Tanaka *et al* 2008). Forkmann and Dangelmayr (1980) showed that, in *Dianthus* flowers, yellow colouration by THC is only possible without chalcone isomerase (CHI) activity. CHI converts THC to the colourless compound (*2S*)-naringenin, which is then further converted into flavones, anthocyanins and other flavonoids. In *Antirrhinum*, a molecule of THC is instead glycosylated by chalcone 4′-*O*-glucosyltransferase (CGT), transported to the vacuole, and converted to aureusidin glucoside by the polyphenol oxidase homologue aureusidin synthase (AS1) (Ono *et al* 2006). This pathway does not interfere with anthocyanin biosynthesis, and orange flowers accumulating both pigments in the same parts of the flower have been observed (Whibley *et al* 2006).

4.1.2.2 Regulation of the aurone biosynthetic pathway by small RNAs

In many *Antirrhinum* species, aurone pigmentation is not present in all parts of the corolla, but rather is restricted to 'foci' – the upper part of the lower lobes, adjacent to the flower opening (**Figure 4.2**). Yellow foci in these species are thought to guide pollinator entry to the flowers (Bradley *et al* 2017). A similar phenotype in *Mimulus lewisii* has been shown to increase pollinator foraging efficiency (Owen and Bradshaw 2011).

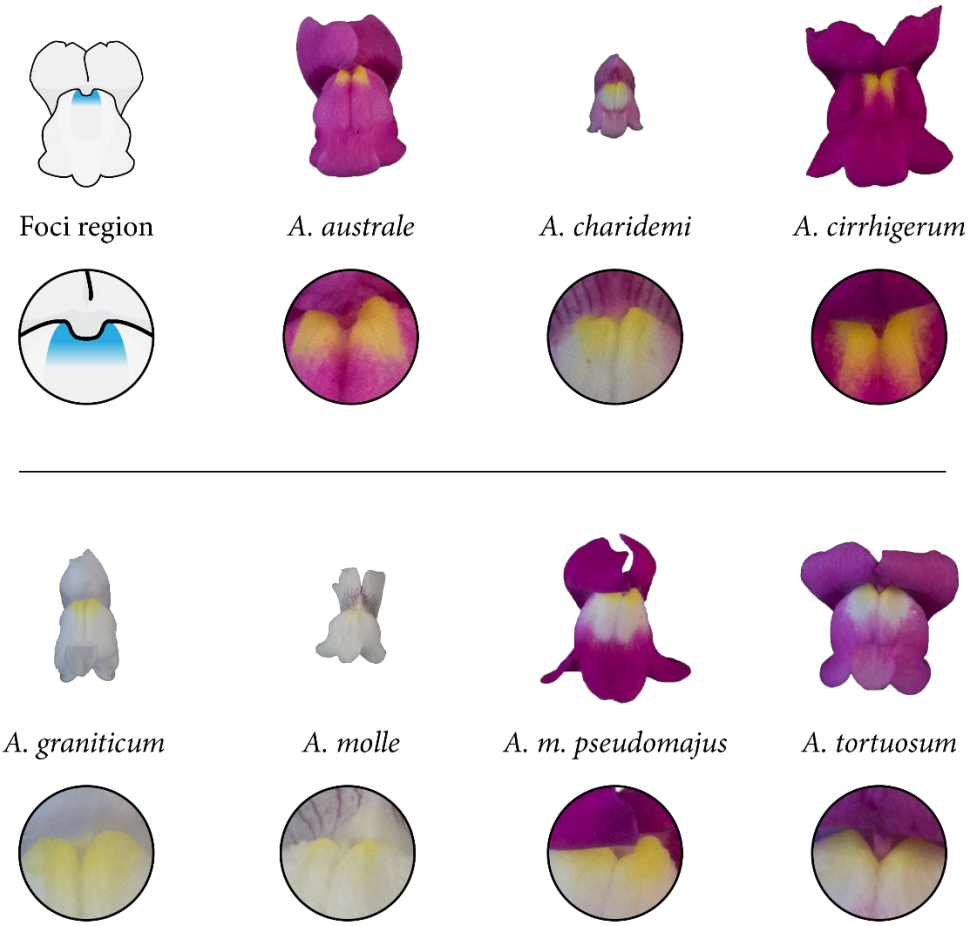


Figure 4.2 Schematic and photographic representation of the region on the flower I refer to as the 'foci'. The diagram on the left shows this region coloured in blue. The foci are found on the lower lobes, adjacent to the opening of the flower. The photographs show the yellow foci found on the flowers of seven example species: *A. australe*, *A. charidemi*, *A. cirrhigerum*, *A. graniticum*, *A. molle*, *A. m. pseudomajus* and *A. tortuosum*. The circled images show the same flowers, but zoomed in on the foci.

In *A. majus* subspecies *pseudomajus*, which has these yellow foci, restriction of aurone pigmentation is regulated by *SULFUREA* (*SULF*), which is located on chromosome 4. *sulf* mutants accumulate yellow pigmentation throughout the flower petals, appearing bright yellow. Bradley *et al* (2017) showed that the *SULF* locus arose through recent inverted duplication of the *FLAVIA* (*FLA*) gene that encodes CGT. The *SULF* locus contains two inverted repeat sequences that generate small RNAs (sRNAs) that repress the *FLA* transcript, thus restricting aurone synthesis. It is not known whether this same mechanism restricts yellow in other *Antirrhinum* species. However, one species with restricted yellow, *A. molle*, is fixed for the recessive *sulf* allele. Its yellow restriction is

thought to arise through changes in the *cis*-regulatory region of the *FLA* gene itself.

4.1.3 Flower colour phenotypes in a hybrid zone between two *Antirrhinum* subspecies

4.1.3.1 *A. m. striatum* and *A. m. pseudomajus*

Two *Antirrhinum majus* subspecies, *A. m. striatum* and *A. m. pseudomajus*, have adjacent population ranges in the Pyrenees, and hybrid zones between the two species have been described (Whibley *et al* 2006). A well-studied hybrid zone between them is discussed in detail in chapter 1. Both subspecies use flower colour to attract bumblebees as pollinators, but their colours and patterns contrast starkly. In *A. m. striatum*, the flowers are yellow with pollinator-guiding magenta veins on the upper lobes. In *A. m. pseudomajus*, the flowers are magenta, with pollinator-guiding yellow foci on the lower lobes surrounded by a white face (**Figure 4.3**).

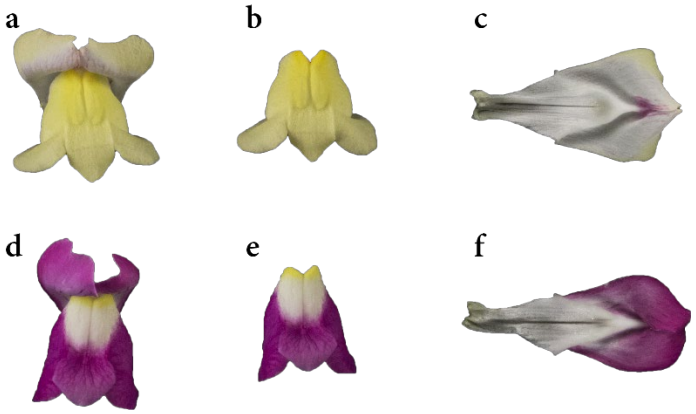


Figure 4.3 Flowers of *A. m. striatum* (a, b and c) and *A. m. pseudomajus* (d, e and f). *A. m. striatum* has full yellow flowers (a). This yellow is particularly bright on the lower lobes (b). On the upper lobes (c), a magenta venation pattern is seen, restricted to the very centre of the flower. *A. m. pseudomajus* has magenta flowers (d), with a white patch in the middle of the lower lobes, and yellow foci (e). Venation pattern, if present, cannot be seen in *A. m. pseudomajus* because the upper lobes are magenta throughout (f).

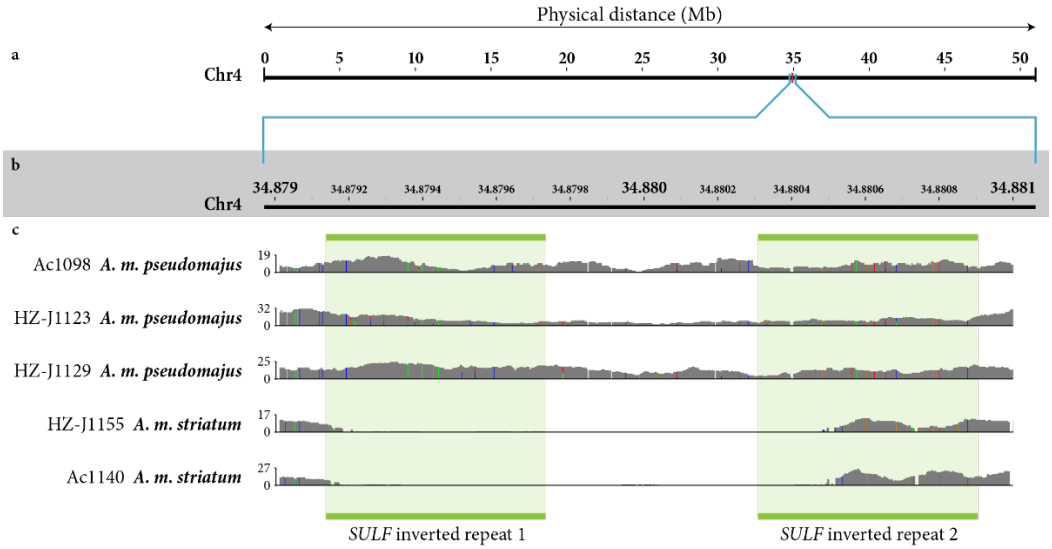


Figure 4.4 Whole genome sequencing depth of coverage along a section of chromosome 4 containing the two *SULF* inverted repeat sequences for three *A. m. pseudomajus* (restricted yellow) and two *A. m. striatum* (spread yellow) individuals. Panel **a** shows the whole length of chromosome 4, with the region expanded in **b** and **c** highlighted. In **c**, the sequenced individuals are shown on the left, and their coverage profiles are illustrated using panels from the Integrative Genomics Viewer (Robinson *et al* 2011), with the *y* axes showing the number of reads mapped at each position in the genome. The locations of the two *SULF* inverted repeats are shown by the green boxes.

The differences in the restriction of yellow pigmentation between *A. m. striatum* and *A. m. pseudomajus* can be explained by genetic differences at the *SULF* locus. Whole genome sequencing of individuals from the two species shows that sequencing depth of coverage is greatly reduced in *A. m. striatum* relative to *A. m. pseudomajus* in a 100-150 kb region that includes *SULF* (Bradley *et al* 2017). Some individuals from *A. m. striatum* have a ~1.4 kb deletion at the *SULF* locus relative to *A. m. pseudomajus*. This deletion covers most of the first inverted repeat and part of the second (**Figure 4.4**).

4.1.3.2 Hybrid phenotypes

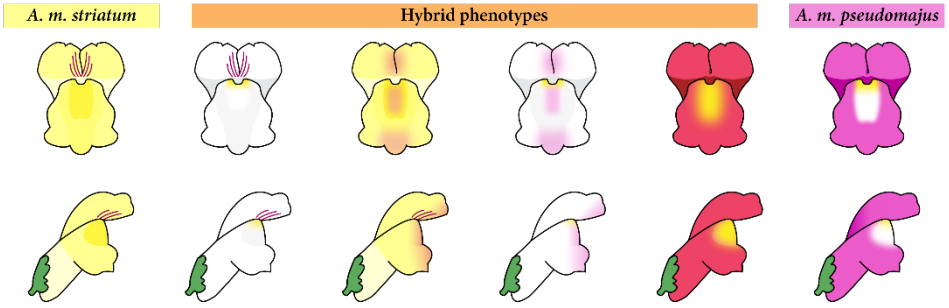


Figure 4.5 Four of the hybrid phenotypes found in the hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. The typical parental subspecies' phenotypes are also shown.

In the kilometre-long core of the hybrid zone, plants with flower colour phenotypes not seen in between *A. m. striatum* and *A. m. pseudomajus* are common. In this area, there are flower colours that are not normally seen in either of the two subspecies outside the hybrid zone. These include white (genotype *SULF ros EL*), pale orange (*sulf ROS EL*), pink (*SULF ROS EL*) and bright orange (*sulf ROS el*) (**Figure 4.5**).

4.1.4 Other yellow variation in the *Antirrhinum* genus

One *Antirrhinum* species is notable for its near lack of yellow. *A. sempervirens* Lapeyr. grows in the western Pyrenees in southwestern France (**Figure 4.6**). This species has white flowers with a very subtle hint of yellow pigmentation around the foci region of the flower, although some accessions appear to have no yellow at all. The flowers also have tightly restricted magenta veins (**Figure 4.7**). *A. sempervirens* has the *SULFUREA* genotype *sulf/sulf* (Bradley, unpublished results), which suggests that the near lack of yellow in this species is regulated by another locus. *A. sempervirens* belongs to subsection Kickxiella, unlike *A. majus*, which belongs to subsection Antirrhinum (Wilson and Hudson 2011). White flowers are characteristic of the Kickxiella group. These plants also have smaller flowers than *A. majus* and tend to grow on rocky cliffs, unlike most subsection Antirrhinum member, which grow on roadsides and other disturbed habitats.

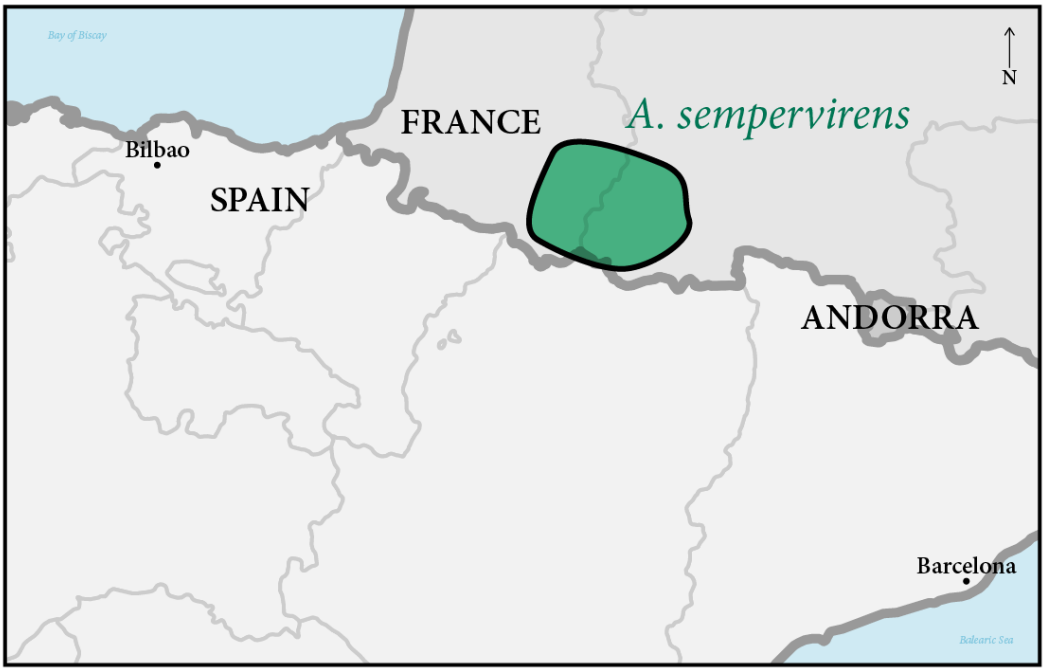


Figure 4.6 Approximate distribution range of *A. sempervirens* near the border between France and Spain. Drawn from information provided in Whibley (2004) and Wilson and Hudson (2011).

Figure 4.7 Images of *Antirrhinum sempervirens* and its flowers. Front view of a flower in the wild (**a**), side view of the same flower in the wild (**b**), growth habit of the plant in the wild (**c**) – all taken in the C-NAP location. Front (**d**) and side (**e**) view flower photographs from of a plant germinated from C-NAP seed. The scale bar (1 cm) applies to **d** and **e**.

Another species, *A. charidemi* Lange., also has an unusual yellow flower colour phenotype compared to the rest of the *Antirrhinum* genus. This species grows in southern Spain (**Figure 4.8**) and has pink flowers with bright yellow foci and yellow pigmentation in the tube of the flower. *A. charidemi* is a member of the Antirrhinum subsection of *Antirrhinum* species (Hudson *et al*, unpublished results), although many

of its phenotypic characters are more typical of the Kickxiella group – its flowers are small (1-1.5 cm in length) and it grows on rocky cliffs – and it has previously been classified as such (Wilson and Hudson 2011). Its flowers accumulate much more yellow pigment in the corolla tube than has been observed in other species.

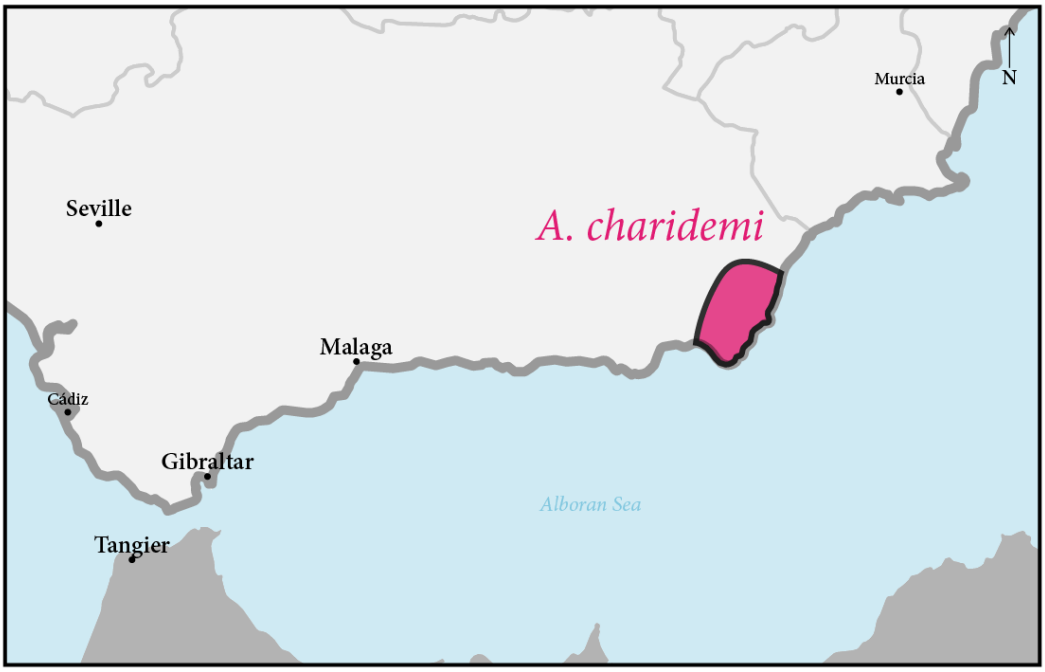


Figure 4.8 Approximate distribution range of *A. charidemi* in southern Spain. Drawn from information provided in Whibley (2004) and Wilson and Hudson (2011).

Figure 4.9 Images of *Antirrhinum charidemi* and its flowers. Front view of a flower in the wild (**a**), side view of another flower in the wild (**b**), growth habit of a plant in the wild (**c**) – all taken in the Y-GAT location. Front (**d**) and side (**e**) view flower photographs from of a plant germinated from Y-GAT seed. The scale bar (1 cm) applies to **d** and **e**.

4.1.5 Using segregating populations to study natural variation

Phenotypic variation between naturally-occurring species and populations can be difficult to work with genetically. While the hybrid zone between *A. m. striatum* and *A. m. pseudomajus* makes it easy to study traits under selection between individuals from those two subspecies in close sympatry, work on species or populations growing in allopatry is more challenging. One way of overcoming this if species are inter-fertile is to cross wild-collected accessions to plants with a known genetic background to generate F2 populations. That way, populations segregating for traits of interest can be generated, allowing their genetics to be studied in a more homogeneous background.

The John Innes Centre maintains an extensive collection of *Antirrhinum* cultivars, as discussed in chapter 2. One of these, line JI7, has been used extensively for research on flower colour. This line is highly inbred, so plants of this variety can be assumed to be homozygous across all loci in the genome (Tavares 2014). Its genome is also sequenced, making bioinformatic analysis of plants from experimental crosses using this cultivar easier and more effective.

4.1.6 Aim of this work

My aim in the experiments described in this chapter is to characterise the variation seen in yellow flower colour in *Antirrhinum* that has not previously been explained and to determine its genetic basis. I will focus on variation from the hybrid zone between *A. m. striatum* and *A. m. pseudomajus* described in Whibley *et al* (2006), as well as the species *A. sempervirens* and *A. charidemi*, both of which have unusual yellow flower colour phenotypes compared to other *Antirrhinum* species. Many MYB-like transcription factors are known to regulate magenta flower colour in *Antirrhinum*, but regulation of yellow colour has so far only been attributed to two loci: *FLA*, a biosynthetic gene involved in aurone production; and *SULF*, which generates regulatory sRNAs to inhibit *FLA*'s function. Magenta anthocyanins and yellow aurones are produced in different parts of the same pathway. Given this, one hypothesis to explain additional yellow variation that has yet to be characterised is that there will be transcription factors involved in its regulation. If this is correct, populations segregating for these yellow colour phenotypes will also segregate for the genes encoding these MYB-like proteins, providing a route for their identification using bulked segregant analysis. This would reveal previously unidentified loci whose

translated sequences contain MYB domains. Alternatively, yellow flower colour may have a different mode of regulation from that of magenta colour, possibly with additional loci encoding regulatory sRNAs or by changes in the *cis*-regulatory regions of one or both biosynthetic genes involved in the aurone pathway. In these instances, allele frequency differences between BSA pools would be seen either at the loci that encode *FLA* or *AS1* themselves or at inverted repeats that show sequence homology to one of these genes, as *SULF* does *FLA*.

4.2 **Results:** Novel phenotypes arise when hybrid zone accessions are crossed with lab cultivars

In an F2 population generated to study flower colour variation in the hybrid zone between *A. m. striatum* and *A. m. pseudomajus*, a previously unseen trait was observed. The corolla of flowers in this population was generally white, but some individuals showed a yellow band in the dorsal petal lobes, just above the opening of the flower (**Figure 4.10**), which we called a 'yellow arc' because of the shape of the pattern.

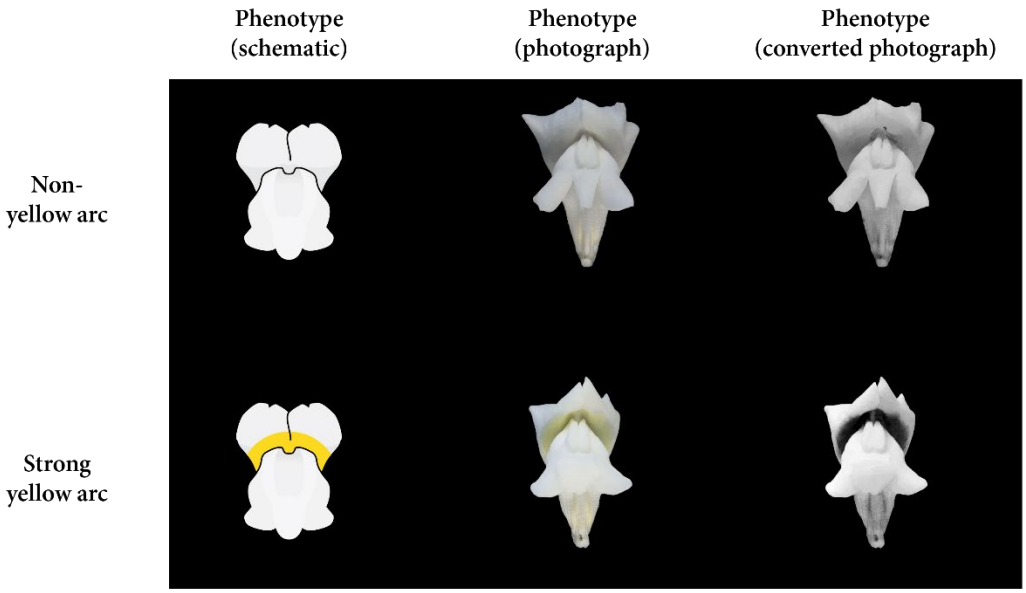


Figure 4.10 The yellow arc phenotype compared with the wildtype non-yellow arc phenotype. From left to right, the first column shows a schematic representation of each of the two phenotypes. The second column shows a photograph of representative flowers, taken from the underside of the flowers. The yellow colour can be difficult to see in photographs. For this reason, I have included a third column, where the photograph from the second column is selectively converted to black and white, with yellow pixels darkened. This conversion makes the yellow arc appear as an easily observed black stripe. The dark patch extending up the tube of the flower also means that this individual has a yellow tube phenotype.

The family where this phenotype was seen was an F2 population from a cross between a hybrid zone-derived plant and a lab cultivar. The male parent of this cross was D194-3, which was germinated from seed collected in the hybrid zone in 2012. It had white flowers without magenta or yellow pigmentation anywhere on the flower lobes, although there was some magenta pigment at the base of the flower tube (**Figure 4.11**). The original wild accession had been sampled to the west of the centre of the hybrid zone, where most surrounding flowers were magenta. It also had predominantly white flowers, but it did have yellow foci and restricted magenta veins (**Figure 4.12**). This flower colour – unlike the colour seen in *A. m. striatum* or *A. m. pseudomajus* outside the hybrid zone – indicates that the plant is of hybrid origin, the result of interbreeding between the two subspecies.

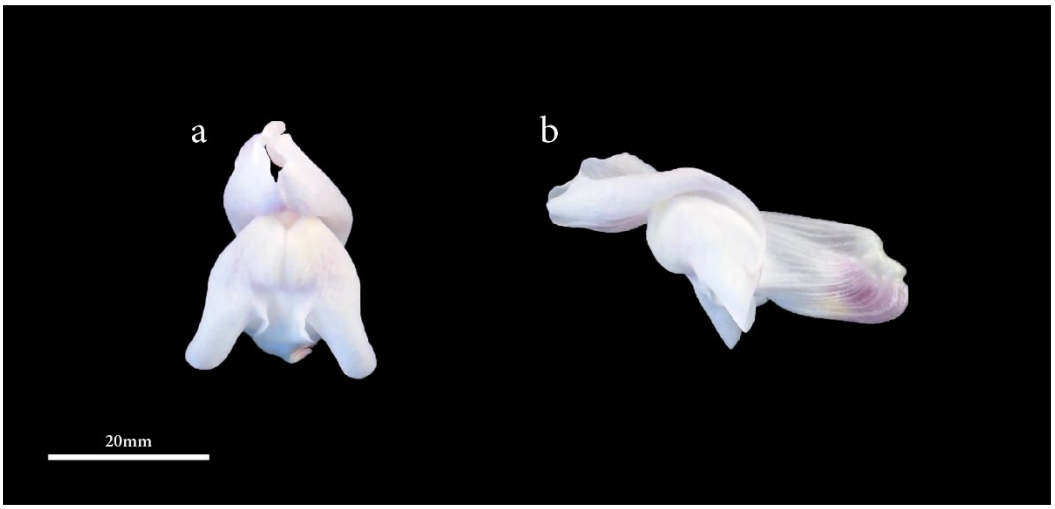


Figure 4.11 Photographs of a flower from D194-3, which grew from a seed collected from M0194, collected in a hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. A front view (a) and a side view (b) are shown.

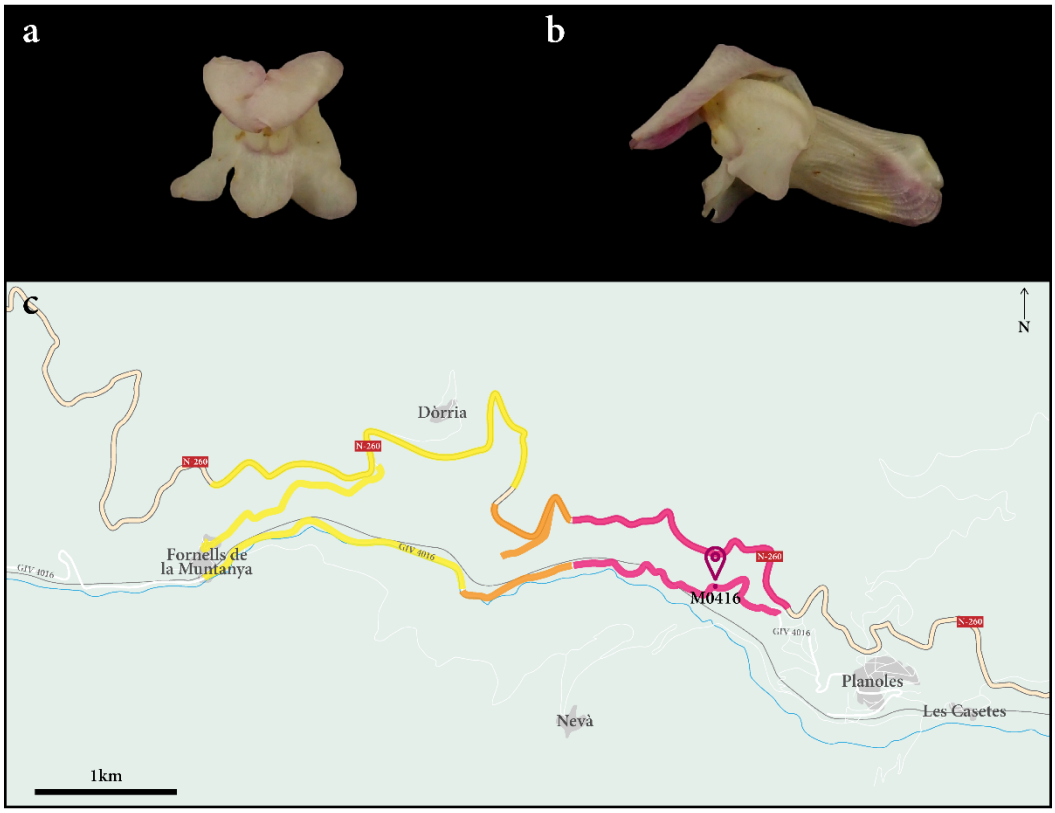


Figure 4.12 Photographs and collection location of accession M0416 collected in a hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. The flower is mostly white (a), although yellow can be seen on the face in the side view (b); this view also reveals the restricted magenta venation pattern on the dorsal lobes, which are hidden because of the photograph's angle in the front view. This plant was sampled within the 'magenta flank' of the hybrid zone – near the hybrid zone itself, but where most of the surrounding flowers have an *A. m. pseudomajus*-like magenta flower colour (c).

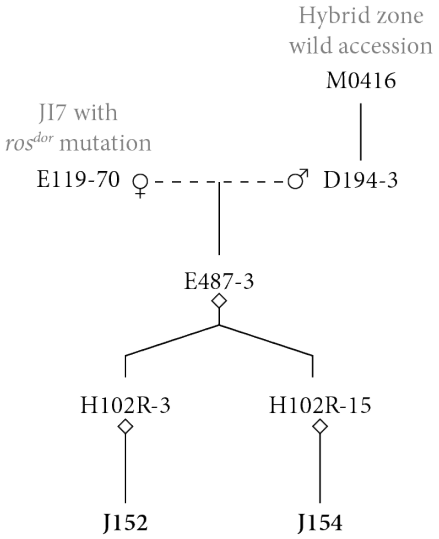


Figure 4.13 Families used for analysing the yellow arc phenotype and their pedigrees. Both families used were F3 populations from a single cross between a plant generated from wild-collected seed and an *A. majus* lab cultivar, but came from two different F2 individuals. M0416 was the wild accession from which the seed that generated D194-3 was collected. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

Originally, the cross between D194-3 and *A. majus* was performed to look at the venation pattern seen on the flowers – D194-3 has very restricted magenta veins (**Figure 4.11 a**). In the F2 that resulted from this cross (H102), however, some of the progeny had a pattern of yellow pigment on the dorsal lobes of the flowers, forming an arc shape above the flowers' foci (**Figure 4.10**). This phenotype had not previously been observed in the wild, and is not seen in D194-3 or in M0416. Two individuals from this F2 population were self-fertilised to make F3 populations, both of which segregated for this yellow arc phenotype. The pedigrees for these families are shown in **Figure 4.13**.

These two small F3 families, comprising 48 plants each were phenotypically scored by Lucy Copsey. In the combined families, around two thirds of individuals had a yellow arc phenotype (**Figure 4.14**). These limited results suggested a 3:1 segregation ratio in these two families for yellow arc and no yellow arc, respectively. A *G*-test for goodness-of-fit gives a *p*-value of 0.244 for J152 and 0.103 for J154, which means that the observed segregation ratio does not differ significantly from a 3:1 ratio expected for

segregation at a single causal locus where one allele is dominant. This gave rise to a hypothesis that the yellow arc phenotype was regulated by a single gene. The lack of observation of the phenotype in the hybrid zone may be because of epistasis (eg the yellow arc may be masked by the full yellow pigmentation of *A. m. striatum*) or a lack of flower colour scoring for the dorsal arc flower region.

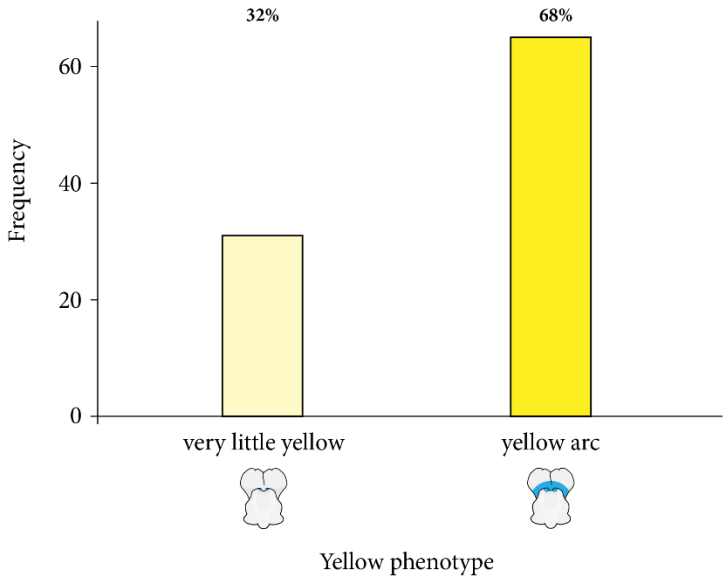


Figure 4.14 Number of individuals scored as having each yellow phenotype in J152 and J154 combined and the proportion of the whole combined family with that phenotype, shown as percentages. Illustrations below the graph columns show schematic representations of each phenotype, with yellow shown in blue to make distinguishing the phenotypes easier.

4.3 **Results:** Bulked segregant analysis and individual genotyping show that the yellow arc phenotype is linked to the *FLAVIA* locus

I attempted to map the yellow arc variation in J152 using bulked segregant analysis. Individuals with the strongest yellow arc pattern and those without any yellow arc pattern were gathered together to construct two bulks with opposing phenotypes. The yellow arc bulk contained 29 individuals and the non-yellow arc bulk contained 10 individuals. DNA from the leaves of these plants was prepared, pooled and sequenced. Mean depth of coverage was calculated as 39× for the yellow arc pool and 41× for the non-yellow arc pool. I mapped the data to the *Antirrhinum* reference genome, processed the data as described in chapter 2 and analysed the resulting data by

calculating allele frequency differences and G' values for each comparison of the yellow arc and non-yellow arc bulks.

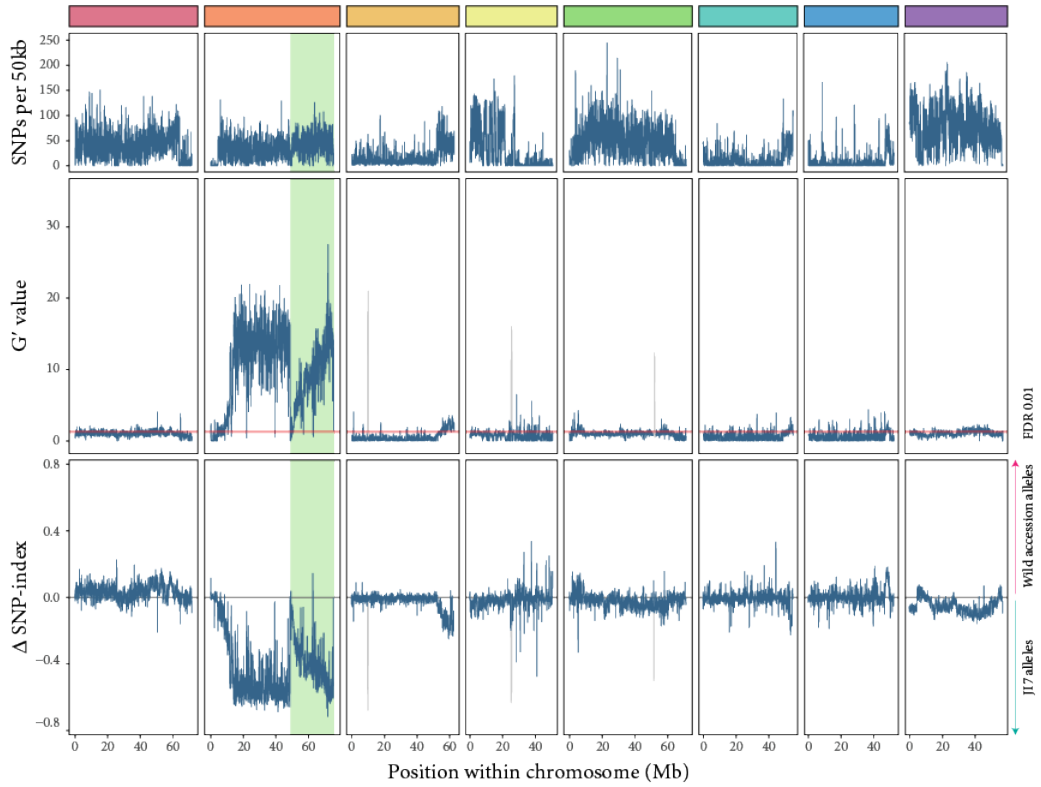


Figure 4.15 Bulked segregant analysis Manhattan plots for family J152 segregating for the yellow arc phenotype. The top plot shows the number of SNPs in each 50 kb window across each chromosome. The middle plot shows the G' value for each SNP. This is a version of the G value averaged across 50 kb windows and smoothed using a tri-cube kernel function. The red line on this plot represents a G' threshold corresponding to a false discovery rate of 0.01. The bottom plot shows the difference in allele frequency of each pool, again averaged and smoothed across 50 kb windows. A negative value indicates that, in a majority of sequencing reads, the JI7 reference genome nucleotide is found at SNPs in that window; a positive value indicates that a majority of reads have a non-reference nucleotide at SNPs in the window.

The plots in **Figure 4.15** show the results from this analysis. The G' values calculated for this family (middle row of plots) are low for most genomic regions, with a high peak covering much of chromosome 2, a low peak at the end of chromosome 3 and a slight elevation on chromosome 8. There are high, narrow peaks seen on chromosomes 3, 4 and 5. Given that these are very narrow and that their heights consistently match parts of chromosome 2, these are likely to be artefacts of genome misassembly in the current version of the *Antirrhinum* genome. These sequences,

should the genome be reassembled, should be investigated and, if appropriate, their positions corrected.

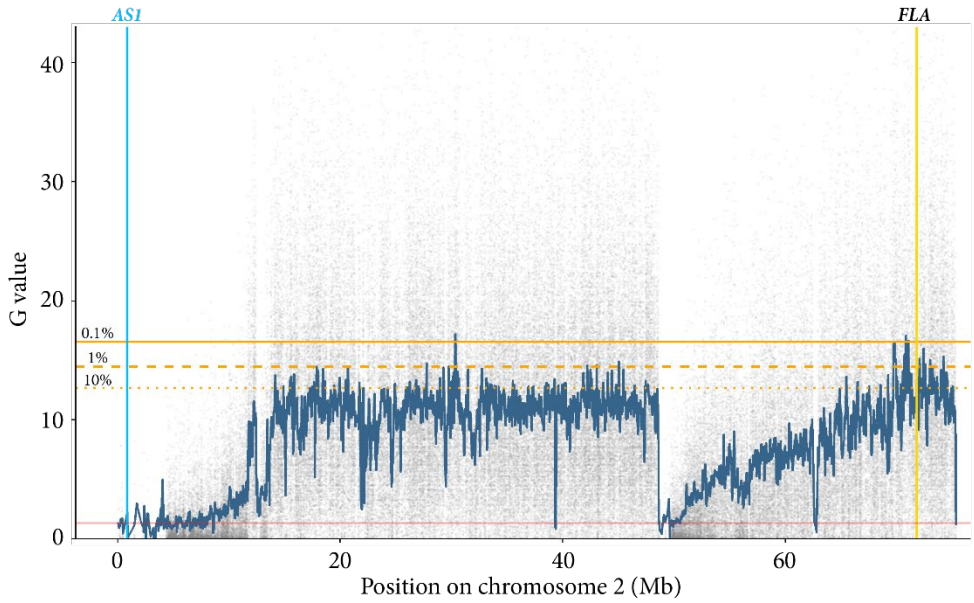


Figure 4.16 A closeup view of chromosome 2 showing *G* values across that chromosome, using the same BSA data as **Figure 4.15**. High *G* values are found across most of the chromosome except for at the start of the chromosome and around the 45Mb position. The grey points in the background show the raw *G* values (*G* calculated for individual SNPs) and the blue line is the *G′* value, calculated using a kernel-adjusted mean value of *G* across 50 kb sliding windows. The positions of two genes found on this chromosome are shown with vertical lines: *AUREUSIDIN SYNTHASE 1* (blue) and *FLAVIA* (gold). The orange horizontal lines represent the top 10% (dotted), 1% (dashed) and 0.1% (solid) thresholds for *G′* on the chromosome. The pale red line corresponds a false discovery rate of 0.1.

As discussed in chapter 3, chromosome 2 contains both of the genes that encode the two enzymes involved in synthesising the yellow pigment aureusidin glycoside from chalcone: *AS1*, which encodes aureusidin synthase, and *FLA*, which encodes chalcone glucosyltransferase (see the aurone biosynthetic pathway in **Figure 4.1** on page 117) (Boell and Bradley, unpublished results). As shown in **Figure 4.16**, *AS1* does not fall within the peak in *G′* value, but *FLA* does. This suggests that mutations at or linked to the *FLA* locus may be involved in establishing the yellow arc phenotype in *Antirrhinum*. Such a mechanism, where biosynthetic genes differ between species to regulate flower colour, would be in contrast with that regulating magenta flower colour, where transcription factors have evolved to interact with various stages of the

anthocyanin biosynthesis pathway.

In the bottom row of plots in **Figure 4.15**, which shows allele frequency differences between the pools across the genome, the peak on chromosome 2 has a negative value. This indicates that the chromosome 2 allele linked to the yellow arc phenotype comes, not from the accession collected in the wild, but from the lab cultivar used in the cross. However, this phenotype is not seen in *A. majus* cultivars that have not been crossed to this wild accession, suggesting that there may be epistatic interactions between an *A. majus* allele at a causal locus on chromosome 2 and an unlinked locus fixed for the wild accession allele.

I also performed the same bulked segregant analysis on J154, the second family that segregated for the yellow arc pattern. As with J152, individuals with the strongest yellow arc pattern and those without any yellow arc pattern were gathered together to construct two bulks with opposing phenotypes. The yellow arc bulk this time contained 16 individuals and the non-yellow arc bulk contained 15 individuals. DNA was collected, prepared and sequenced as for J152. Mean depth of coverage was calculated as 58× for the yellow arc pool, but was considerably lower – at 23× – for the non-yellow arc pool.

Genome-wide G' values calculated for 50 kb sliding windows across each chromosome are shown in **Figure 4.17**. As with the plot for J152, the largest peak is seen on chromosome 2 and includes the *FLA* locus, which is consistent with *FLA* or another gene on the same chromosome being responsible for the yellow arc phenotype. The lack of recombination on chromosome 2 again makes it impossible to determine using this data where exactly on the chromosome the linked locus is located. This analysis also confirms that chromosome 2 in individuals with a strong yellow arc mostly carries alleles from the *A. majus* cultivar at most SNPs, suggesting that the yellow arc phenotype comes from this research line rather than from the wild accession. However, unlike in the J152 plot shown in **Figure 4.15**, J152 shows additional peaks in G' and allele frequency difference (Δ SNP-index) on chromosomes 4 and 5 and the end of chromosome 1. These additional peaks may mean that there may be several unlinked loci contributing to the phenotype. The wide nature of these peaks also suggests a lack of recombination as seen on chromosome 2. However, the poorer sequencing coverage in one of the bulks means that the non-yellow arc bulk is under-represented compared to the yellow arc bulk. This is likely to lead to an increased level of noise in the BSA results, and some of the peaks seen may be artefacts caused by this

noise. This may also explain why the G' and Δ SNP-index lines in **Figure 4.17** are more erratic than those in **Figure 4.15** – with poor coverage, the effect of very small signals can be amplified.

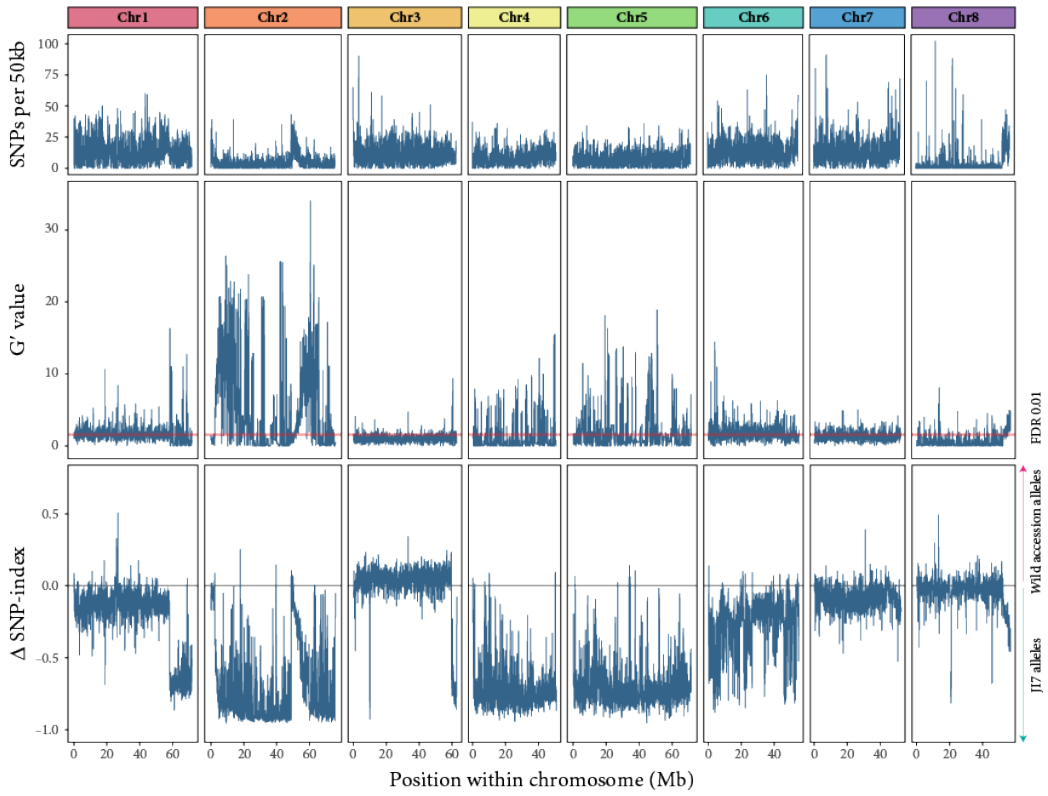


Figure 4.17 Bulked segregant analysis Manhattan plots for family J154 segregating for the yellow arc phenotype. The top plot shows the number of SNPs in each 50 kb window across each chromosome. The middle plot shows the G' value for each SNP. This is a version of the G value averaged across 50 kb windows and smoothed using a tri-cube kernel function. The red line on this plot represents a G' threshold corresponding to a false discovery rate of 0.01. The bottom plot shows the difference in allele frequency of each pool, again averaged and smoothed across 50 kb windows.

Using a larger window size gives a cleaner G' signal because it filters out noise – high-frequency deviations in G which have nothing to do with the trait being analysed. This can prove useful, especially when analysing bulks from a population such as this one, where the parent plants that gave rise to the population used were not closely related and the number of individuals in each bulk was relatively small. To look for the regions of the genome most strongly associated with the arc colour trait, I repeated my analysis using a window size of 1 Mb (**Figure 4.18**).

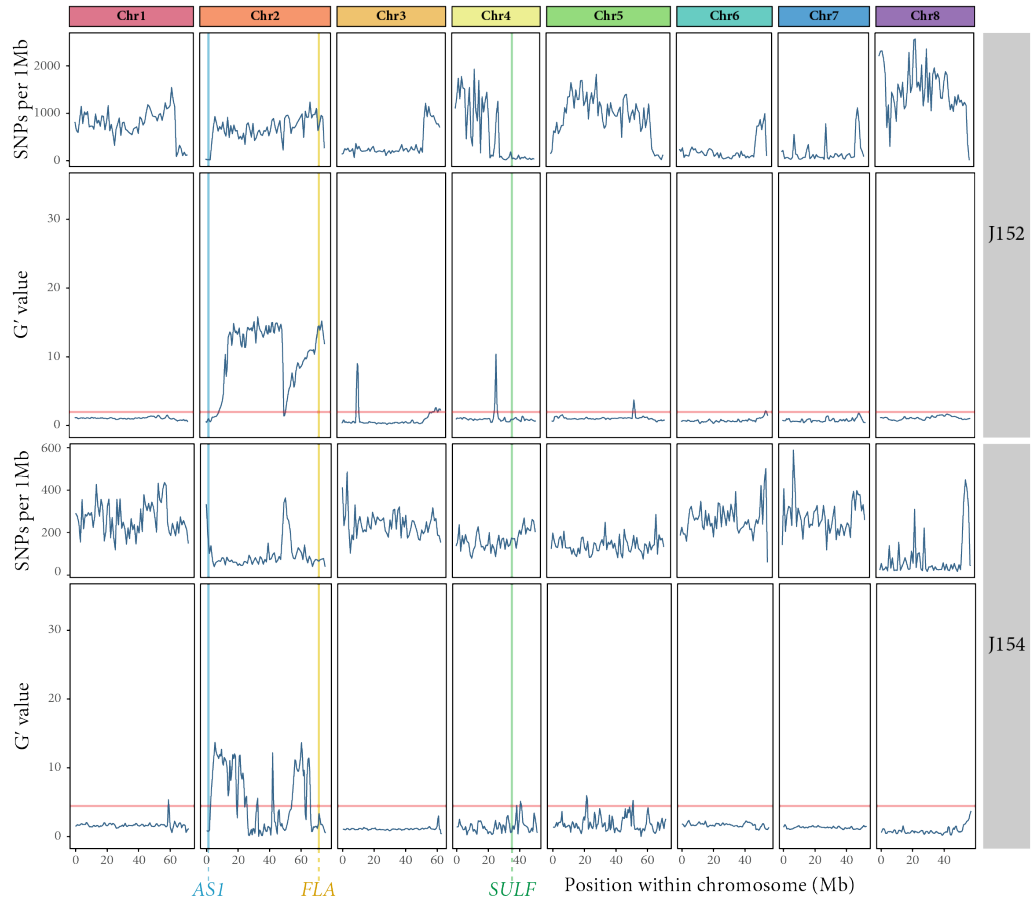


Figure 4.18 Bulked segregant analysis Manhattan plots for family J152 (upper two plots) and J154 (lower two plots) – repeating the analyses shown in **Figure 4.15** and **Figure 4.17**, respectively, but using larger windows of 1 Mb. The top plot of each pair shows the number of SNPs in each window. The bottom plot of each pair shows the G' value for each SNP, averaged across nearby SNPs in a 1 Mb window and smoothed using a tri-cube kernel function. Note that the scale of the *y*-axes in the two plots showing the number of SNPs in each window are different because of the lower SNP density in J154. The positions of the *AS1* and *FLA* genes involved in the biosynthesis of yellow pigmentation are shown with blue and yellow lines, respectively, and the position of *SULF*, known for its regulation of yellow pigmentation, is shown with a green line.

Comparing the G' plot for J152 in **Figure 4.18** with its 50 kb window counterpart in **Figure 4.15**, there is little or no difference in the locations of the peaks seen, and major allele frequency differences are still observed across chromosome 2. The profiles of the narrow peaks on chromosomes 3, 4 and 5 – starting and ending very suddenly – are further evidence that they represent sequences that have been misassembled in the *Antirrhinum* genome. In the J154 analysis, however, using 1 Mb windows has a more notable effect, as seen when comparing **Figure 4.18** with **Figure 4.17**. After increasing

the window size, the height of the peaks on chromosomes 4 and 5 become much lower, suggesting that the signal seen is, at least in part, noise. Such noise can be the result of sequencing coverage not being uniform across the genome. Indeed, this appears to be the case in both J152 and J154, as seen in the SNPs-per-window plots in **Figure 4.18**. Large areas of the genome have very low SNP coverage, such as the first 40-50Mb of chromosome 6 in J152.

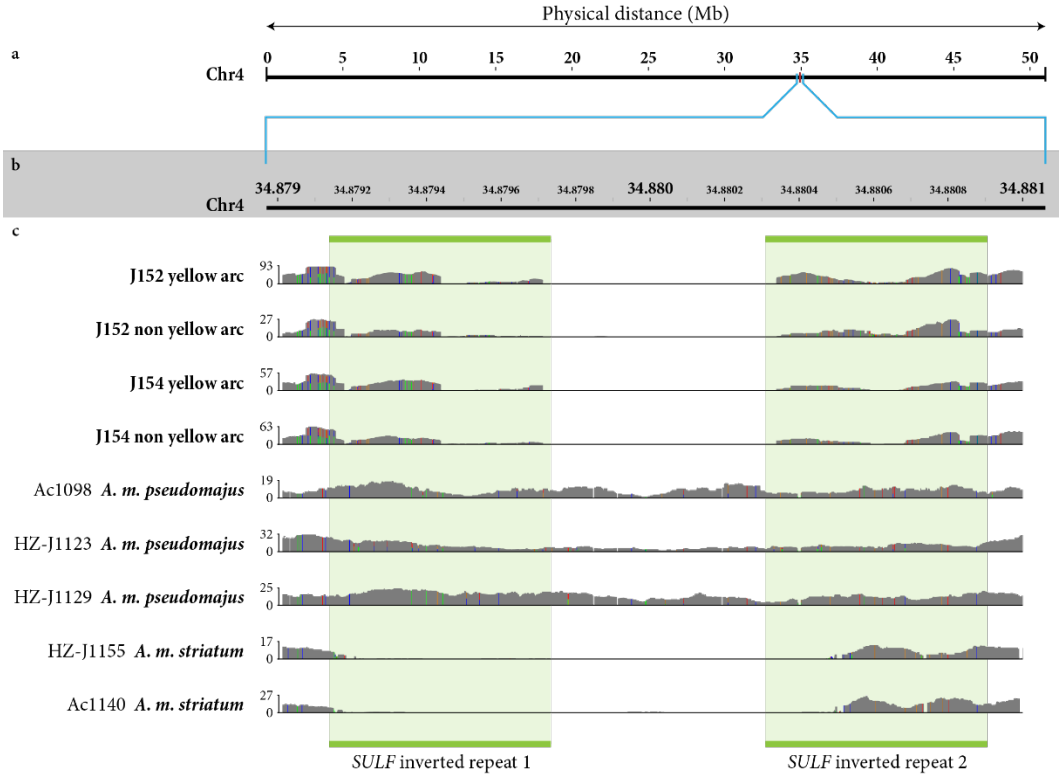


Figure 4.19 Whole genome sequencing depth of coverage along a section of chromosome 4 containing the two *SULF* inverted repeat sequences for samples from J152, J154, three *A. m. pseudomajus* (restricted yellow) and two *A. m. striatum* (spread yellow) individuals. Panel **a** shows the whole length of chromosome 4, with the region expanded in **b** and **c** highlighted. In **c**, the sequenced individuals are shown on the left, and their coverage profiles are illustrated using panels from the Integrative Genomics Viewer (Robinson *et al* 2011), with the *y* axes showing the number of reads mapped at each position in the genome. Coloured vertical lines show positions that contain SNPs relative to the reference genome. The locations of the two *SULF* inverted repeats are shown by the green boxes.

A key regulator of yellow pigmentation in *A. m. pseudomajus* is the *SULF* gene, described in section 4.1.2.2. Small RNAs transcribed at this locus inhibit the expression of *FLA*, thus limiting the amount of yellow pigmentation produced in specific tissues. In *A. m. pseudomajus*, this results in yellow pigmentation being limited

to the flower foci. I looked at the mapped sequencing reads from J152 and J154 in the Integrative Genomics Viewer (Robinson *et al* 2011) and compared the *SULF* locus of these families to the same position in *A. m. pseudomajus* and *A. m. striatum* (**Figure 4.19**). Both J152 and J154 appear to have partial deletions relative to *A. m. pseudomajus* covering parts of the *SULF* inverted repeat and the sequence linking them. This is a smaller deletion than in *A. m. striatum*. This deletion is fixed in both phenotypic bulks for both families, which explains why no peak is seen at this locus in the BSA plots.

The results at *SULF* suggest that yellow arc is an allele of *SULF* (*SULF*^{arc}) that is present at an unknown frequency in the hybrid zone, but that its phenotype is not visible when combined with the *A. m. pseudomajus* allele of *FLA* (*FLA*^{P}). When crossed with a line that has the *A. majus* JI7 allele of *FLA* (*FLA*^{7}), which has a similar *FLA* sequence to the brightly yellow-flowered *A. m. striatum*, the *SULFarc* pattern becomes visible. Thus, different combinations of *FLA* and *SULF* alleles in the hybrid zone can give rise to a range of flower colour phenotypes (**Figure 4.20**). Compared with the *A. m. pseudomajus* allele of *SULF* (*SULF*^{P}), the *SULF*^{arc} allele appears to be a weak inhibitor of yellow pigmentation. Rather than limiting yellow pigmentation to the flower foci, *SULF*^{arc} restricts aurone production to a broader region of the flower that includes the dorsal arc. This weakened restriction effect may be because fewer sRNAs are produced,

owing to the partial deletion.

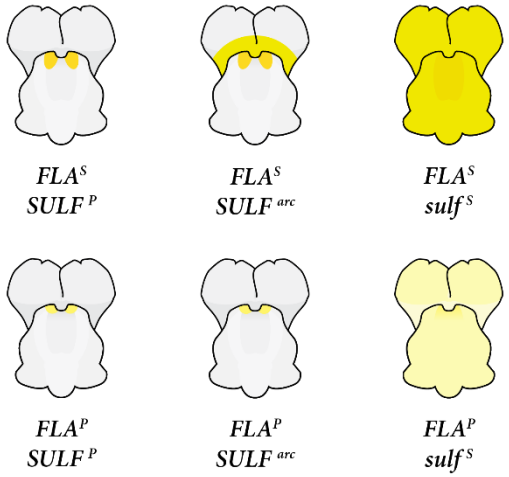


Figure 4.20 Hypothesised phenotypic effect of different combinations of *FLA* and *SULF* alleles in the hybrid zone between *A. m. striatum* and *A. m. pseudomajus*. A superscript *S* represents an allele from *A. m. striatum* and *P* represents an allele from *A. m. pseudomajus*. *SULF arc* represents the yellow arc allele of *SULF* found in the hybrid zone.

4.4 **Results:** Genotyping for *FLAVIA* reveals close linkage between genotype and phenotype

J152, the family used to generate the BSA results presented in section 4.3 was small, with only 46 individuals. To confirm the effect of the *A. majus* JI7 allele of *FLAVIA* (*FLA*$^{?}$) on the yellow arc phenotype, seeds from the same crosses that generated J152 and J154 (same parent but different capsules) were sown to generate two larger families, L122 and L123 (see **Figure 4.13**), with 200 and 160 individuals, respectively. These segregated for the yellow arc pattern in a similar way to J152 and J154. Some flowers in L122 and L123 also showed a yellow pigmentation in the tube of the flower, another phenotype not usually seen in wild accessions. Further analysis of photographs from J152 and J154 shows that some individuals these families also had

yellow tubes.

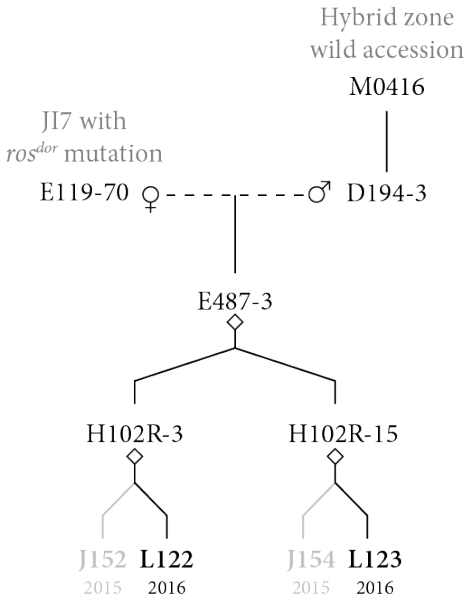


Figure 4.21 Families used to confirm the *FLA* genotypes of the yellow arc phenotype and their pedigrees. Both families used were F3 populations from a single cross between a plant generated from wild-collected seed and an *A. majus* lab cultivar, but came from two different F2 individuals. L122 and L123 came from different selfed flowers (ie separate capsules) on the same plants as J152 and J154, respectively. M0416 was the wild accession from which the seed that generated D194-3 was collected. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

I photographed a sample flower from each individual in L122 and L123 and scored their flower colours based on these photographs. Of the 360 individuals in the combined population, 54 were not in flower when I scored them, giving a total phenotyped population size of 306. I photographed each flower from two angles: from the underside of the flower, where the front of the dorsal petals, the flower face and the underside of the flower tube were visible; and from the top of the flower, where the back of the dorsal petals and the top of the tube were visible. To make flower colour scoring easier and more accurate, I selectively converted the flower photographs to black and white, making the yellow pigments darker than those of other colours, using an automated processing script in Adobe Photoshop, as shown previously in **Figure**

4.10.

Of the 306 individuals scored in L122 and L123, 214 individuals (69.9%) had a yellow arc phenotype, while the remaining 92 individuals (30.1%) had no yellow arc. However, these families showed additional variation not scored in J152 and J154. Of the plants showing a yellow arc pattern, 68 individuals (31.7% of those with a yellow arc, 22.2% of the family) had yellow pigmentation in the tube of the flower. 146 individuals (68.2% of those with a yellow arc and 47.7% of the family) had a yellow arc pattern without yellow pigmentation in the tube (**Figure 4.22**). A *G* test for goodness of fit showed that this ratio did not differ significantly from a 1:2:1 ratio (p = 0.06), suggesting that the *A. majus* allele responsible for the increased yellow production is semidominant.

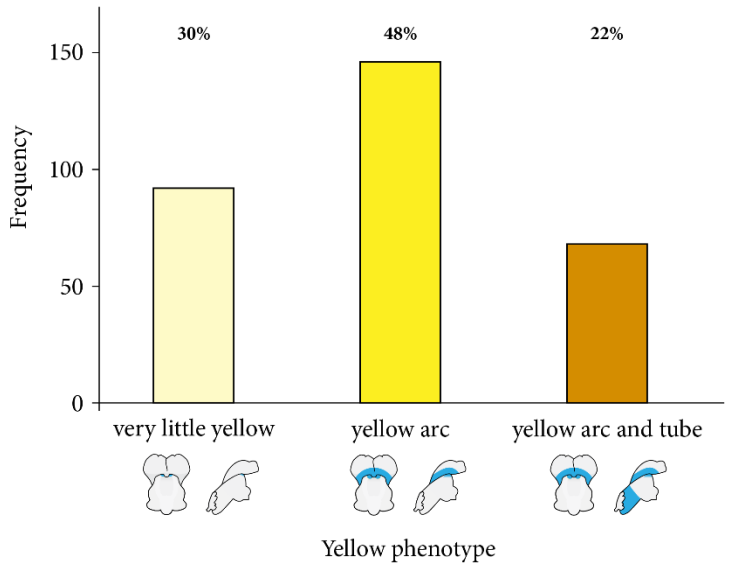


Figure 4.22 Number of individuals scored as having each yellow phenotype in L122 and L123 combined and the proportion of the whole combined family with that phenotype, shown as percentages. Illustrations below the graph columns show schematic representations of each phenotype, with yellow shown in blue to make distinguishing the phenotypes easier.

A hypothesis to explain the arc and tube phenotypes based on this segregation ratio would be that, in the presence of *SULF* ^{arc}, one copy of *FLA*^{7} gives a yellow arc phenotype, while an extra copy (ie *FLA*^{7}/*FLA*^{7}) additionally gives a yellow tube phenotype (hypothesis 1). Alternatively, the yellow tube phenotype may be regulated by an unlinked gene that is epistatic to *FLA*, requiring the *FLA*^{7} allele for its phenotype

to show (hypothesis 2). If hypothesis 1 is correct, plants with both a yellow arc and a yellow tube phenotype will have the genotype FLA^7/FLA^7, while those with only a yellow arc phenotype will have the genotype FLA^7/FLA^P. If hypothesis 2 is correct, all plants with a yellow arc, regardless of tube colour, will have the genotypes FLA^7/FLA^7 or FLA^7/FLA^P; plants with yellow flower tubes will not be distinguishable by their *FLA* genotypes.

I designed seven sets of Kompetitive Allele-Specific PCR (KASP) oligonucleotide primers (LGC Ltd 2013) at and near the *FLA* coding region on chromosome 2 to determine the genotypes of individuals in L122 and L123. I also used five oligonucleotide primers developed by LGC Ltd for the same region, albeit based on a closely related population from the hybrid zone rather than L122 and L123 themselves. The positions of all these primers on chromosome 2 in the reference *Antirrhinum* genome are shown in **Table 4.1** and in **Figure 4.23**. Of the primers I designed, none revealed polymorphisms in the individuals I tested, but one of the LGC-designed primer pairs did. I used this – primer pair 4, with a focal SNP at 71,833,823 bp in the promoter region of *FLA* – to genotype the combined families.

Table 4.1 KASP oligonucleotide primer pairs designed for determining the genotypes of individual plants at and near the *FLA* coding region on chromosome 2, along with the result of testing the primers on a test plate containing 96 individuals with a selection of different yellow phenotypes. Although referred to as pairs, three primers are used in each analysis: two in the same direction, each ending in a different allele of the focal SNP; and one common primer in the opposite direction. Manually designed primers were the ones I designed based on genome sequencing data.

Primer pair	Focal SNP (bp)	Origin	Testplate result
1	71832425	Manually designed	Monomorphic
2	71833304	LGC-designed	Monomorphic
3	71833653	Manually designed	Monomorphic
4	71833823	LGC-designed	Polymorphic
5	71834117	Manually designed	Monomorphic
6	71834290	Manually designed	Monomorphic
7	71834360	Manually designed	Monomorphic
8	71834537	LGC-designed	Monomorphic
9	71835496	Manually designed	Monomorphic
10	71836472	Manually designed	Monomorphic
11	71852521	LGC-designed	Monomorphic
12	71872855	LGC-designed	Monomorphic

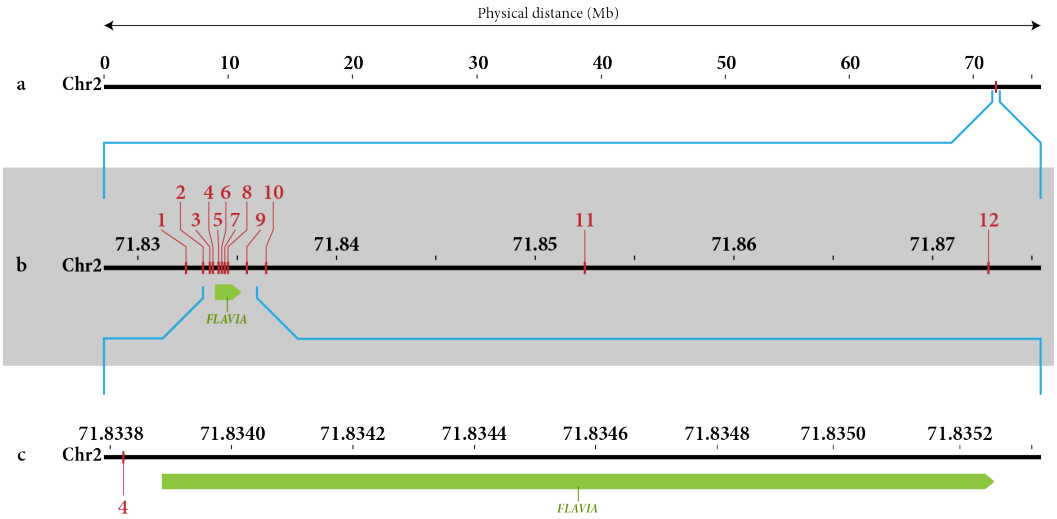


Figure 4.23 Positions of the focal SNPs (red lines) of the primers described in **Table 4.1**, relative to the coding region of *FLA* (green block); a shows the position of the region containing *FLA* on chromosome 2; b shows the location of *FLA* and the 12 focal SNPs more specifically; and c shows the position of the focal SNP of primer pair 4, used for subsequent analyses to genotype for *FLA*.

Of the 337 reactions whose genotypes I could confidently call, 75 wells (22.3%) only showed fluorescence corresponding to the JI7 allele (FLA^{7}), 97 wells (28.9%) only showed fluorescence corresponding to the alternative, *A. m. pseudomajus* allele (FLA^{P}), and the remaining 165 wells (49.0%) showed fluorescence in both channels, meaning that both alleles are present (ie heterozygous, FLA^{7}/FLA^{P}). These values showed the expected 1:2:1 ratio for two segregating alleles, as confirmed by a *G* test for goodness of fit, which revealed that the expected and observed ratios did not differ significantly ($p = 0.113$).

I then looked at the association between yellow phenotype and *FLA* genotype in the population (**Table 4.2**). Of the 60 individuals homozygous for FLA^{7} for which I also had phenotype information, all individuals had a yellow arc, and all but three individuals (95.0%) had a yellow tube. Every one of the 85 phenotyped individuals homozygous for FLA^{P} lacked the yellow arc pattern. Of the 148 individuals that were heterozygous at *FLA* for which I also had phenotype information, 22 individuals (14.9%) had no yellow arc, 121 individuals (81.8%) had a yellow arc without yellow pigmentation in the tube, and five (3.4%) had a yellow tube as well as a yellow arc.

Table 4.2 Frequencies of different phenotypes given the *FLA* genotypes of plants in L122 and L123. The phenotypes are: no yellow arc, NY; yellow arc, YA; and yellow arc and tube, YT.

Genotype	FLA^7/FLA^7			FLA^7/FLA^P			FLA^P/FLA^P		
Phenotype	NY	YA	YT	NY	YA	YT	NY	YA	YT
Number	0	3	57	20	121	5	85	0	0

These results, with some yellow tube plants heterozygous at *FLA*, appeared to confirm hypothesis 2 – that the yellow tube phenotype is regulated by a separate gene from *FLA* at an unlinked locus. However, given that only three out of 57 FLA^7 homozygotes did not have a yellow tube and only five out of 121 FLA^7/FLA^P heterozygotes did have a yellow tube, I could not exclude the possibility that hypothesis 1 was correct and that some of my samples were misgenotyped or mislabelled.

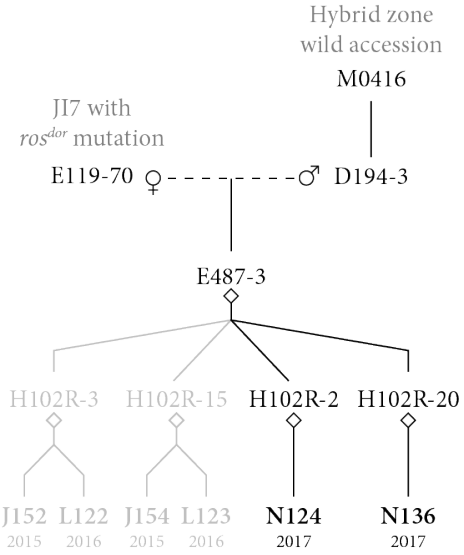


Figure 4.24 Families used to confirm the *FLA* genotypes of the yellow arc phenotype and their pedigrees. Both families used were F3 populations from a single cross between a plant generated from wild-collected seed and an *A. majus* lab cultivar, but came from two different F2 individuals. L122 and L123 came from different selfed flowers on the same plants as J152 and J154, respectively. M0416 was the wild accession from which the seed that generated D194-3 was collected. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

I used two further plant populations grown in 2017 – N124 and N136 – to verify my

results from L122 and L123. These F3 families were derived from different individuals in the same F2 that gave rise to J152, J154, L122 and L123 (**Figure 4.24**). A total of 167 plants in N124 and 223 plants in N136 were phenotyped for yellow flower colour and genotyped for *FLA*. The ratios of phenotypes in these families differed substantially from those seen in the previous families (**Figure 4.25**). In N124, 24 individuals (14.4%) had little or no yellow pigmentation in the flowers, and the other 143 (85.6%) had strong yellow pigmentation at the flower foci. Of these 143 with strong yellow pigmentation, 80 had a yellow arc (55.9% of those with strong yellow pigmentation, 47.9% of the family) and 63 did not. And of those with a yellow arc, 53 also had a yellow tube (66.3% of those with a yellow arc, 31.7% of the family), while the remaining 27 did not. In N136, 47 individuals (21.1%) had little or no yellow pigmentation in the flowers, and the other 176 (78.9%) had strong yellow pigmentation at the flower foci. Of these 176 with strong yellow pigmentation, 142 had a yellow arc (80.7% of those with strong yellow pigmentation, 63.7% of the family) and 34 did not. And of those with a yellow arc, 34 also had a yellow tube (23.9% of those with a yellow arc, 15.3% of the family), while the remaining 49 did not.

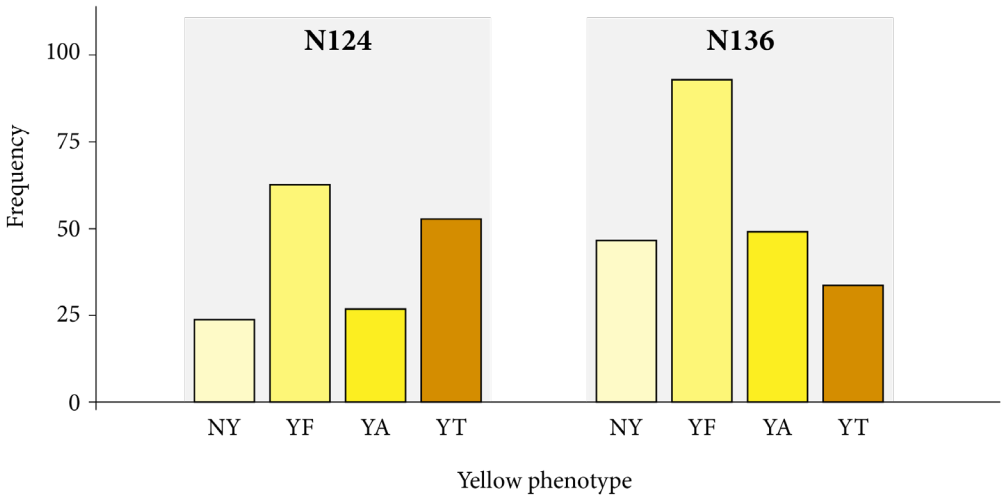


Figure 4.25 Frequencies of different yellow phenotypes in N124 and N136: little or no yellow (NY); yellow only on the foci (YF); yellow on the foci and yellow arc (YA); and yellow in the foci, arc region and flower tube (YT).

Looking at the genotypes, in N124, 53 individuals (31.7%) were homozygous for FLA^{7}, 38 (22.8%) were homozygous for FLA^{P}, and 76 (45.5%) were heterozygous. In N136, 46 individuals (20.6%) were homozygous for FLA^{7}, 51 (22.9%) were homozygous for FLA^{P}, and 126 (56.5%) were heterozygous. In both families, a *G* test

for goodness of fit did not show significant differences between the observed ratio and the 1:2:1 expected ratio for a segregating marker with two alleles.

Table 4.3 Linkage between *FLA* genotype and yellow phenotype in N124 and N136, using the following phenotypic categories: little or no yellow (NY); yellow only on the foci (YF); yellow on the foci and yellow arc (YA); and yellow in the foci, arc region and flower tube (YT).

Genotype	FLA^7/FLA^7				FLA^7/FLA^P				FLA^P/FLA^P			
Phenotype	NY	YF	YA	YT	NY	YF	YA	YT	NY	YF	YA	YT
N124 Frequency	0	0	0	53	0	49	27	0	24	14	0	0
N136 Frequency	0	0	12	34	0	89	37	0	47	4	0	0

There was a strong association between the genotype at *FLA* and the yellow phenotype scored for each individual in these two families (**Table 4.3**). In N124, all 53 of those with a FLA^7/FLA^7 genotype had yellow pigmentation in the foci, arc and tube. None of the 38 plants with the FLA^P/FLA^P genotype had a yellow arc or a yellow tube, although 14 had strong yellow pigmentation at the foci while the rest did not. The 76 heterozygotes had moderate yellow phenotypes – at least strong yellow on the face, with 27 also showing a yellow arc phenotype but none with yellow tubes. In N136, 34 of the 46 plants with a FLA^7/FLA^7 genotype had yellow pigmentation in the foci, arc and tube, but the remaining 12 lacked the yellow tube phenotype. As in N124, none of the 51 plants with the FLA^P/FLA^P genotype had a yellow arc or a yellow tube, and only four had strong yellow pigmentation at the foci, with the other 47 showing little or no yellow in the flowers. Heterozygotes, again, had either strong yellow foci or yellow foci and yellow arcs, and none had yellow tubes.

These results confirm that the yellow arc phenotype requires at least one copy of the FLA^7 allele and appear to suggest that the yellow tube phenotype is only seen in FLA^7/FLA^7 homozygotes, although this contradicts the results seen in L122/L123. However, not all FLA^7 homozygotes in N136 have a yellow tube, which means that I cannot infer that my earlier hypothesis 1 – that yellow tube arises from having two copies of FLA^7 – is correct. It is possible, therefore, that there is another locus unlinked to *FLA* that regulates yellow pigmentation in the flower tube, but that it is epistatic to *FLA* (tube hypothesis 1, **Table 4.4**). A future experiment to identify this locus would be to pool the 12 N136 individuals with a FLA^7/FLA^7 genotype and a yellow arc but no yellow tube into one bulk, and individuals with the same genotype but with a yellow tube into another bulk, and to sequence them for bulked segregant analysis. A better experiment

would be to re-sow selfed seed from H102R-20, the parent of N136, to get larger numbers for these pools. The genotyping results for N124 and N136 also confirm that plants homozygous for FLA^{P} do not show a yellow arc phenotype, although they may accumulate yellow pigment in the flower face. An alternative hypothesis is that the variation in yellow tube in N136 plants fixed for FLA^{7} is seen because of environmental differences between individual plants (tube hypothesis 2, **Table 4.4**). If this is correct, BSA of the pools described above would show no peaks.

Table 4.4 Predicted outcome, given two alternate hypotheses, of BSA comparing plants from N124/N136 fixed for FLA^{7} but with different tube colour phenotypes.

	Tube hypothesis 1	Tube hypothesis 2
Description of hypothesis	Yellow tube is regulated by an unidentified gene unlinked to *FLA*.	Yellow tube is the result of variation in yellow pigment accumulation because of environmental differences.
Predicted result of BSA of FLA^{7}/FLA^{7} plants with and without yellow tube phenotypes	Peak on a chromosome other than chromosome 2.	No peaks seen.

The different phenotypic ratios in N124/N136 compared to L122/L123 may be explained by the N-set families (and their parents) being differentially fixed for another regulator of yellow pigmentation compared to the L-set families. Because N124 and N136 (or any other families from H102R-2 or H102R-20) have not been sequenced, their *SULF* genotype is unknown, but if the *SULF* deletion is different in between N124/N136 and L122/L123, the interaction with FLA^{7} may produce different results (arc hypothesis 1, **Table 4.5**). This could be tested through individual Sanger sequencing at the *SULF* locus for individuals from each of the four families. Alternatively, there may be another locus segregating in N124 and N136 that leads to additional variation in yellow colour. This may be the same locus that leads to a yellow tube phenotype when plants have the FLA^{7}/FLA^{7} genotype, although this would not explain the differences in ratios between N124 and N136 (arc hypothesis 2, **Table 4.5**). A way to test this would be to pool *FLA* heterozygotes with and without a yellow arc phenotype and look for BSA peaks outside chromosome 2.

Table 4.5 Predicted outcome, given two alternate hypotheses, of two proposed future experiments.

	Arc hypothesis 1	Arc hypothesis 2
Description of hypothesis	There are fewer individuals with yellow arc because N124/N136 have a different *SULF* deletion compared with L122/L123.	There are fewer individuals with yellow arc because N124/N136 are segregating for a second locus unlinked to *FLA* that changes yellow colour amount/intensity.
Predicted result of Sanger sequencing at *SULF* for N124/N136 and L122/L123	Sequences differ between the N-set families and the L-set families – eg size of the deletion is different.	Sequences are the same for all individuals tested.
Predicted result of BSA of FLA^{7}/FLA^{P} plants with and without yellow arc phenotypes	No peaks seen.	Peak on a chromosome other than chromosome 2.
Predicted result of growing self seed from other H102 individuals.	Not all families show the same yellow arc phenotypic ratios.	Not all families show the same yellow arc phenotypic ratios.

4.5 **Results:** Additional variation in yellow pigmentation is also linked to the *FLAVIA* locus

In addition to the yellow variation in yellow arc seen in the families previously described in this chapter, I looked at variation in other yellow phenotypes from other *Antirrhinum* species. *A. sempervirens* grows in southwestern France and has white flowers with a very subtle hint of yellow in the face region and tightly restricted magenta veins (**Figure 4.26 a** and **b**). *A. charidemi* grows in southeastern Spain and has pink flowers with bright yellow foci and, unlike other species, a yellow flower tube (**Figure 4.26 c** and **d**).

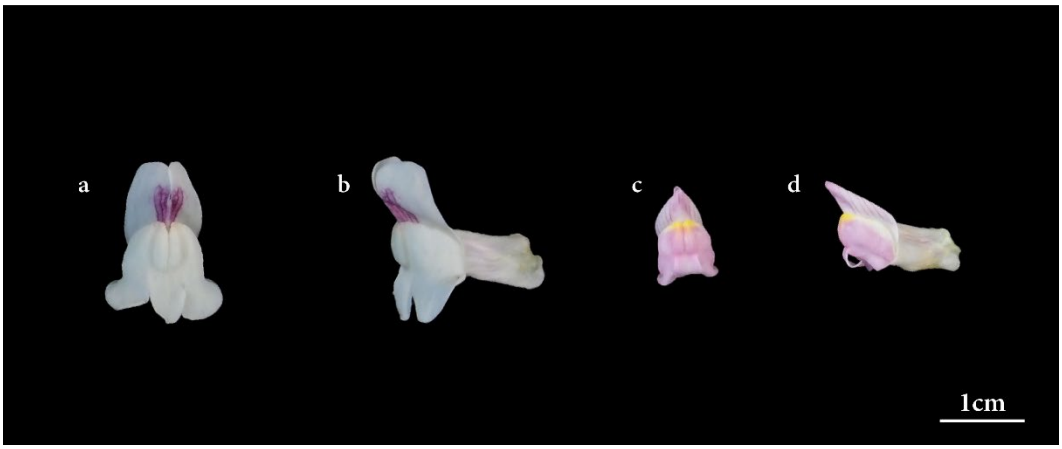


Figure 4.26 Flowers of *A. sempervirens* (**a** and **b**) and *A. charidemi* (**c** and **d**). The scale bar of 1 cm relates to all four images.

Figure 4.27 Location within France of the C-NAP collection location where *A. sempervirens* was sampled in 2003. The location was named after the nearby Pont Napoleon, a bridge in the commune of Luz-Saint-Sauveur, Hautes-Pyrénées department, southwestern France.

Seeds were collected from *A. sempervirens* in the Parc national des Pyrénées in southwestern France in 2003 (**Figure 4.27**). This accession is named C-NAP (C was the identifier given to 2003 as a collecting year and NAP refers to the nearby Pont Napoleon). The seeds were germinated in glasshouse conditions and one of these individuals was crossed with *A. majus* (JI7) and self-fertilised to give an F2 generation. Seeds were also collected from *A. charidemi* in the Cabo de Gata-Níjar natural park in southeastern Spain in 1999 (**Figure 4.28**). This accession was named Y-GAT (Y refers to 1999 as a collection year and GAT is short for Cabo de Gata). Progeny from these,

were recurrently backcrossed to JI7 to introgress alleles of interest. This backcrossed line was then crossed to JI7 to generate an F2 population.

Figure 4.28 Location within Spain of the Y-GAT collection location where *A. charidemi* was sampled in 1999. The location was named after the Cabo de Gata-Níjar Natural Park, eastern Andalucia, where the seeds were collected.

A family was generated each from two crosses between *A. sempervirens* and JI7: H115 and H118. These were scored for magenta and yellow pigmentation. In H115 (**Figure 4.29**), of the 99 plants scored, 18 had no yellow pigmentation, 64 had yellow pigmentation restricted to the flower face and 17 had yellow spread throughout the petal lobes (**Figure 4.30**). Yellow pigmentation in this family did not appear to be regulated by one gene with a semidominant allele; this would mean a 1:2:1 segregation ratio, but a *G* test for goodness of fit showed that the observed ratio differed significantly from this ($p = 0.007$). A likely explanation for the segregation ratio seen is that the family segregated for *SULF* and for another gene regulating yellow pigmentation. The *sulf* allele gives a spread yellow phenotype (like that seen in **Figure 4.30 c** and **f**), but the dominant *SULF* allele, fixed in JI7, only restricts yellow (**Figure 4.30 b** and **e**) and does not eliminate it or weaken its intensity (as is the case in **Figure 4.30 a** and **d**). Therefore, there may be an additional gene, '*NOYELLOW*', regulating yellow pigmentation in H115. If the lack of yellow is associated with a recessive *noyellow* allele, a 9:3:4 ratio of restricted yellow (*SULF*/- *NOYELLOW*/-) to spread yellow (*sulf*/*sulf NOYELLOW*/-) to lack of yellow (-/- *noyellow*/*noyellow*) would be expected because *SULF* would likely be epistatic to *NOYELLOW*. A *G* test for goodness of fit showed that the observed ratio did not differ significantly from this

expected 9:3:4 ratio (p = 0.093). This unknown '*NOYELLOW*' gene was therefore of interest, as this could be a previously unidentified gene responsible for colour variation between *Antirrhinum* species. H115 also segregated for restriction of magenta and the presence/absence of veins (**Figure 4.30**). These magenta phenotypes were not considered in this study, but are likely to be because of segregation of *ELUTA* (which restricts magenta pigmentation in *A. m. pseudomajus* and is thought to serve the same function in *A. sempervirens*) and *VENOSA* (which regulates anthocyanin production in tissue overlying veins in the dorsal lobes of *A. m. striatum* and is thought to serve the same function in *A. sempervirens*).

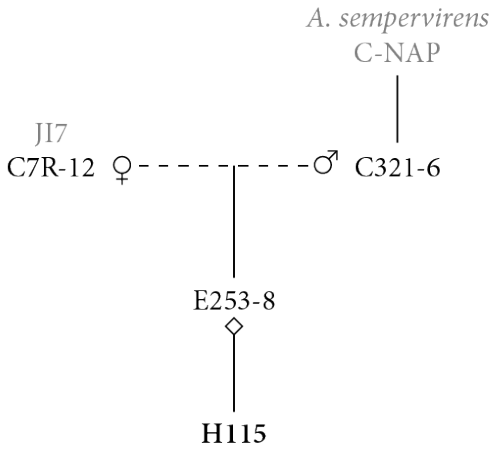


Figure 4.29 The pedigree of H115, which segregated for lack of strong yellow pigmentation on the flower face, as seen in *A. sempervirens*. In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

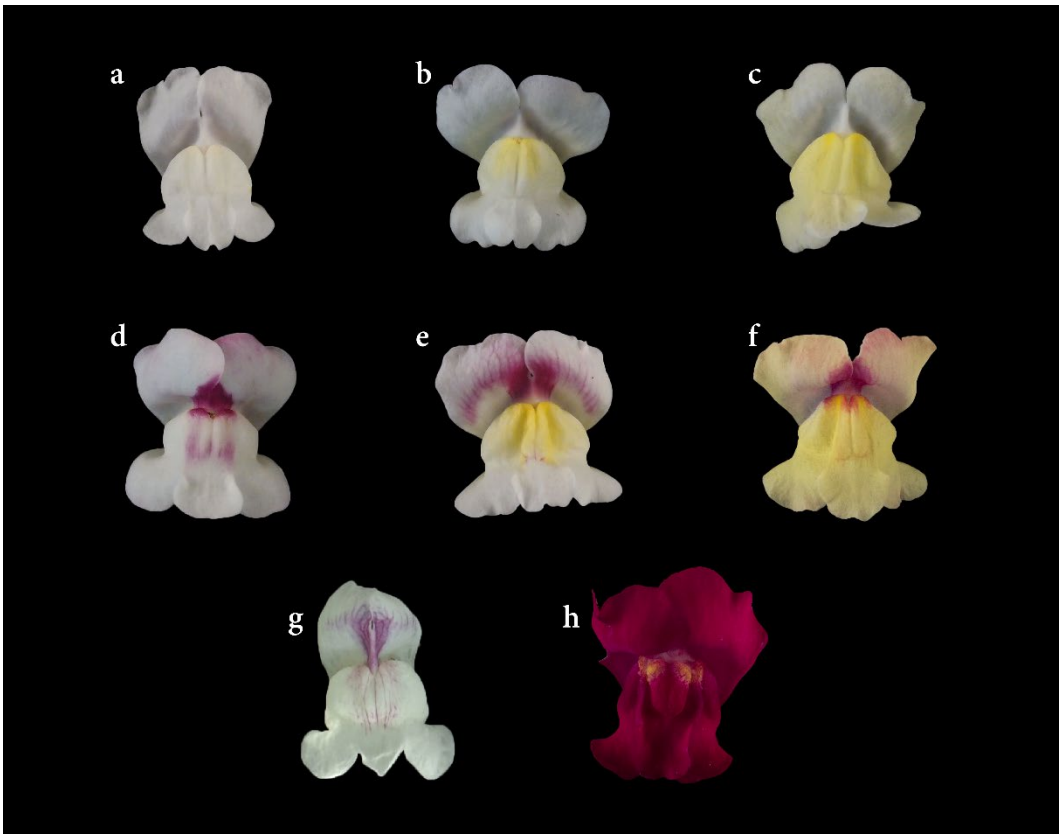


Figure 4.30 Flower colour phenotypes in H115 (**a-f**), an F2 population between *A. sempervirens* (**g**), which has very little yellow pigmentation, and the lab cultivar JI7 (**h**), which has strong but restricted yellow. Many individuals had no yellow pigmentation (**a** and **d**), while others had yellow on the flower face (**b** and **e**). Some were *sulf* mutants, meaning that yellow pigmentation was produced throughout the petal lobes (**c** and **f**). The *A. sempervirens* accession used for this cross has a magenta pattern restricted to the centre of the flower, magenta veins and no yellow pigmentation (**g**). *A. majus* var. JI7 has full magenta pigmentation and restricted yellow foci (**h**).

In H118 (**Figure 4.31**), of the 134 plants scored, 23 had no yellow pigmentation, 82 had yellow pigmentation restricted to the flower face and 20 had yellow spread throughout the petal lobes (**Figure 4.32**). This ratio, like in H115, was not significantly different to the expected 9:3:4 ratio of restricted yellow to spread yellow to lack of yellow ($p = 0.092$). This appears to confirm that yellow flower colour in this family is regulated at *SULF* and at another locus where the recessive allele is associated with *A. sempervirens*'s lack of yellow phenotype.

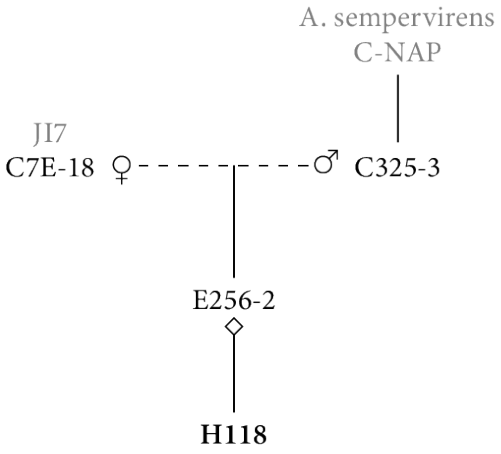


Figure 4.31 The pedigree of H118, which segregated for lack of strong yellow pigmentation on the flower face, as seen in *A. sempervirens.* In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

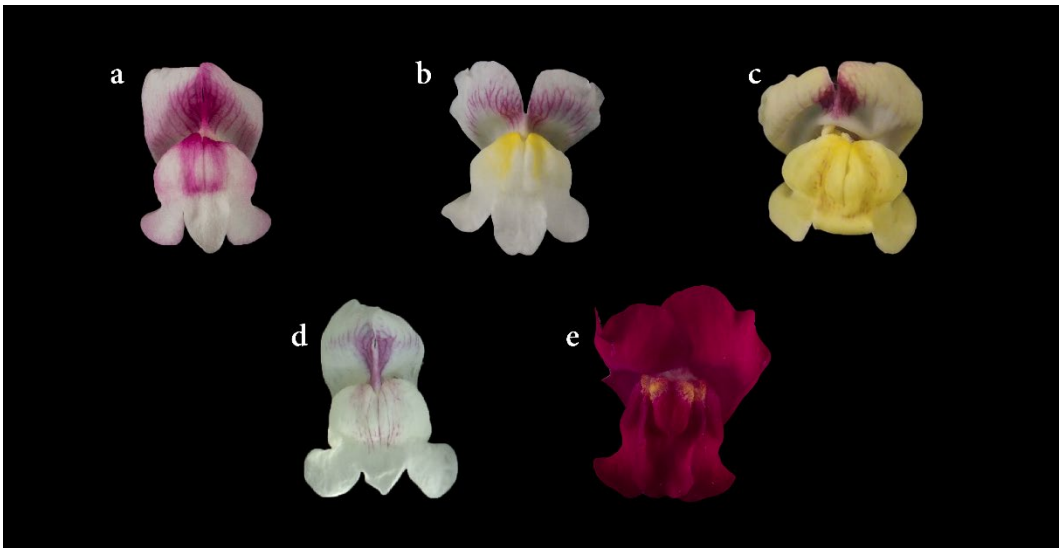


Figure 4.32 Flower colour phenotypes in H118 (**a-c**), an F2 population between *A. sempervirens* (**d**), which has very little yellow pigmentation, and the lab cultivar JI7 (**e**), which has strong but restricted yellow. The yellow colour phenotypes seen were no yellow (**a**), restricted yellow foci (**b**) and spread '*sulf*' yellow (**c**). Variation was also seen in magenta colouration with some showing a *ROSEA-ELUTA* phenotype (**a**), some showing a *VENOSA-eluta* phenotype (**b**) and some showing a *VENOSA-ELUTA* phenotype (**c**). Magenta pigmentation and its regulation is discussed in the introduction to chapter 3. The *A. sempervirens* accession used for this cross has a magenta pattern restricted to the centre of the flower, magenta veins and no yellow pigmentation (**g**). JI7 has full magenta pigmentation and restricted yellow foci (**h**).

H246, the F2 from the cross between an *A. charidemi*-derived family and JI7, segregated for yellow pigmentation in the tube of the flowers (**Figure 4.33**), a phenotype seen in *A. charidemi*. The *A. charidemi*-derived male parent of this cross was the result of several back-crosses to JI7 (**Figure 4.34**), so this family was more introgressed than those used elsewhere in this chapter. There were 67 plants in this family; 16 plants had yellow flower tubes, while 51 did not. This represents a 3:1 ratio between the no yellow tube and strong yellow tube phenotypes, confirmed by a *G* test for goodness of fit (p = 0.832). This suggests that yellow tube is associated with a recessive allele at a single locus.

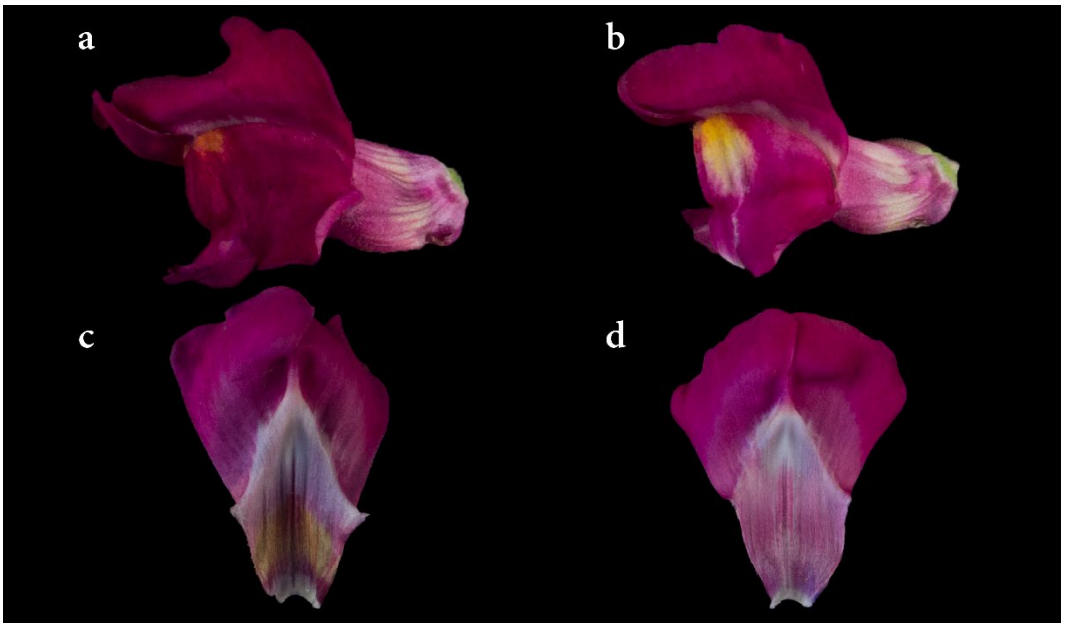


Figure 4.33 Flowers from H246 segregating for the yellow tube phenotype. Some plants had strong yellow pigmentation on the flower tube (**a**), while others lacked this phenotype (**b**). This phenotype was easier to see on the inside of the tube where less magenta colour accumulates (**c** and **d**).

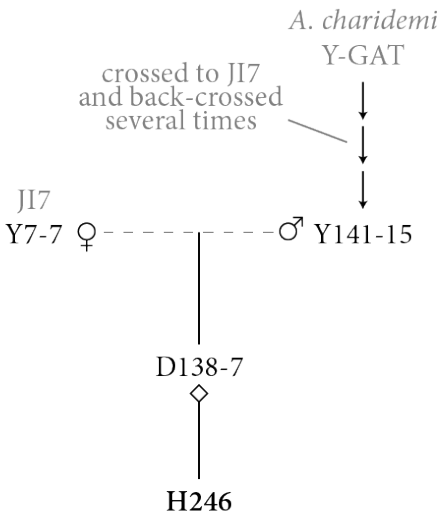


Figure 4.34 The pedigree of H246, which segregated for strong yellow pigmentation in the flower tube, as seen in *A. charidemi.* In the diagram, female and male parents are indicated using their respective symbols (♀ and ♂) and a diamond (◇) represents self-fertilisation. Solid lines show the relationship between parent and progeny and dashed lines show crosses between parents.

Leaves collected from these families were pooled as follows: for H115, one pool of plants with no yellow (18 individuals), one pool with restricted yellow (17 individuals),

and one pool with spread yellow (15 individuals) to confirm that *SULF* was segregating; for H118, one pool of plants with no yellow (20 individuals), one pool with restricted yellow (20 individuals), and one pool with spread yellow (15 individuals); and for H246, one pool of plants with a strong yellow tube phenotype (16 individuals) and one pool with no yellow tube (20 individuals). DNA was extracted from these pooled leaves and sequenced at The Genome Analysis Centre (now the Earlham Institute). I processed and analysed the sequencing data using the same pipeline as for J152 and J154 earlier in this chapter and mapped the variation in yellow pigmentation using bulked segregant analysis.

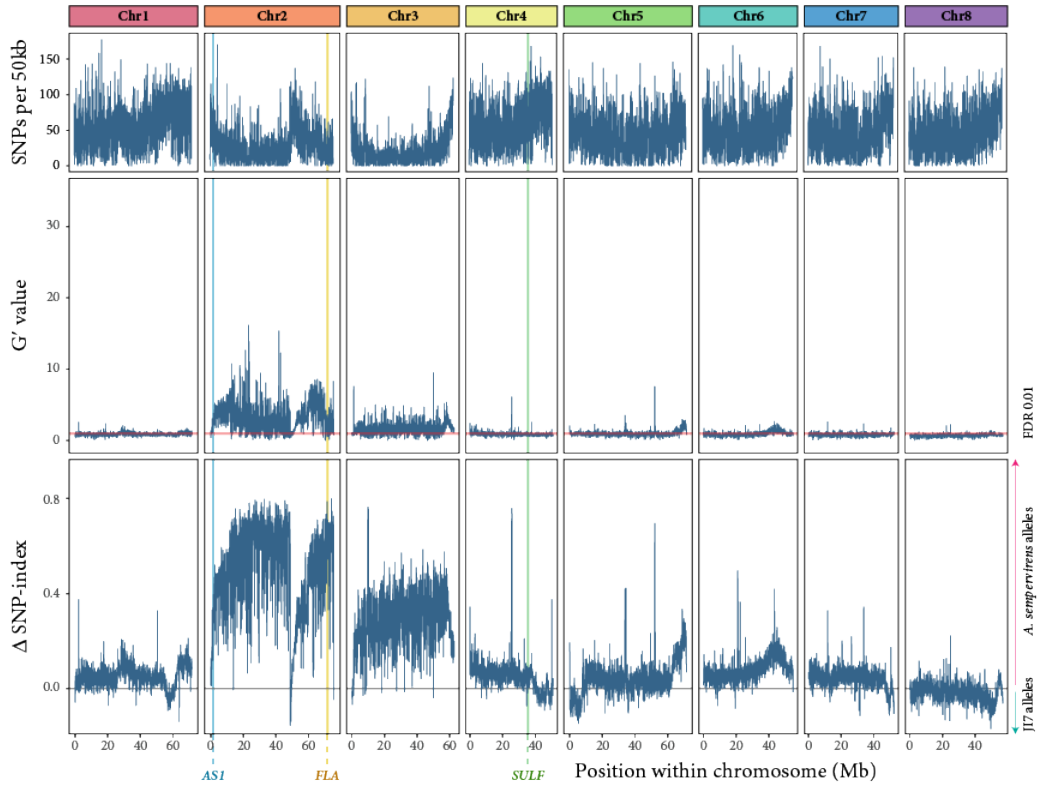


Figure 4.35 Bulked segregant analysis Manhattan plots for family H115 segregating for the presence and absence of yellow pigmentation in the flowers. The top plot shows the number of SNPs in each 50 kb window across each chromosome. The middle plot shows the G' value for each SNP – a modified G value averaged across 50 kb windows and smoothed using a tri-cube kernel function. The red line on this plot represents a G' threshold corresponding to a false discovery rate of 0.01. The bottom plot shows the difference in allele frequency of each pool, again averaged and smoothed across 50 kb windows. A negative value indicates that, in a majority of sequencing reads, the JI7 reference genome nucleotide is found at SNPs in that window; a positive value indicates that a majority of reads have a non-reference nucleotide at SNPs in the window.

The results from the first bulked segregant analysis, comparing the pools with and without yellow from H115, are shown in **Figure 4.35**. The highest peak in G' for this comparison is on chromosome 2, with most windows along the chromosome having G' values above the false discovery rate. Elevated G' is also seen along much of chromosome 3. Finally, there is a very low peak at the end of chromosome 5, and another one towards the end of chromosome 6.

The highest peak in G' is on chromosome 2, and the Δ SNP-index shows that *A. sempervirens* alleles are more common than the JI7 allele in the no-yellow pool on this chromosome. This suggests that a gene on this chromosome is associated with the lack of yellow phenotype from *A. sempervirens*. The peak on chromosome 2, as with previous work on the yellow arc phenotype, is wide because a lack of recombination is seen on this chromosome between JI7 and *Antirrhinum* species collected in the wild.

Both *AS1* and *FLA*, the genes that encode the two enzymes that convert chalcone to the yellow aurone pigment, are located on chromosome 2. However, the highest points of the G' and Δ SNP-index peaks exclude *AS1*. *FLA* is included in the highest peak in Δ SNP-index, although it, too, has a lower G' value than most regions on the genome. This could be because SNP density is slightly lower around the *FLA* locus compared to some other parts of the chromosome. However, there may be an additional regulator of yellow flower colour located on this chromosome that has not previously been identified.

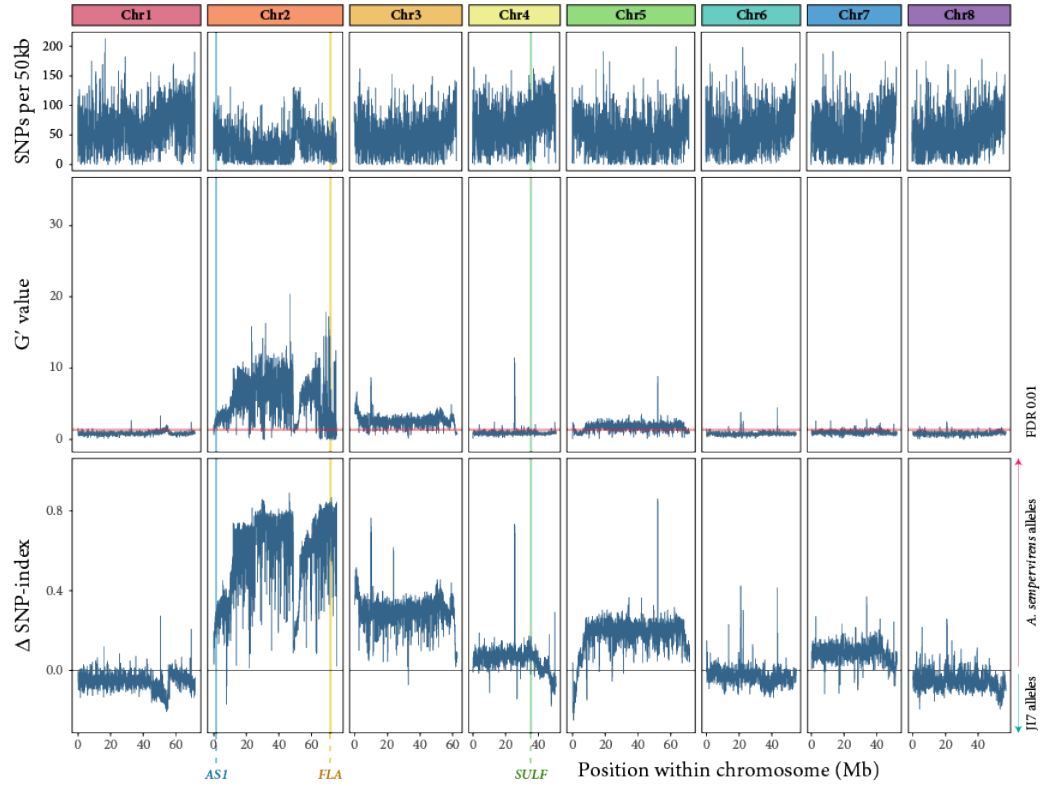


Figure 4.36 Bulked segregant analysis Manhattan plots for family H118 segregating for the presence and absence of yellow pigmentation in the flowers. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools. All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA* and *SULF* are indicated with vertical lines and labelled below the *x* axis.

Bulked segregant analysis of H118 largely confirmed the result seen for H115 (**Figure 4.36**). The highest peak, again, is on chromosome 2 and excludes *AS1*. As seen in H115, however, there is a slight dip in G' value around the *FLA* locus, although this is less prominent in H118. No such dip is seen in Δ SNP-index at the same locus.

I also used bulked segregant analysis to compare plants in these families that had restricted yellow pigmentation with those that had spread yellow pigmentation to confirm that the phenotype was regulated by *SULF* as was predicted. In H115, the largest peak was on chromosome 4 and includes the *SULF* locus (**Figure 4.37**). The same result was seen for H118 (**Figure 4.38**). There was a small signal around 60Mb along chromosome 1 and another around 50Mb along chromosome 2 in H118.

However, these are much lower than the peak on chromosome 4. Because there were very few individuals in these pools (20 and 15), a difference in a few individual plants may produce a large signal on the plot. This may also be true for the low peaks in the bulked segregant analysis of the lack of yellow phenotype.

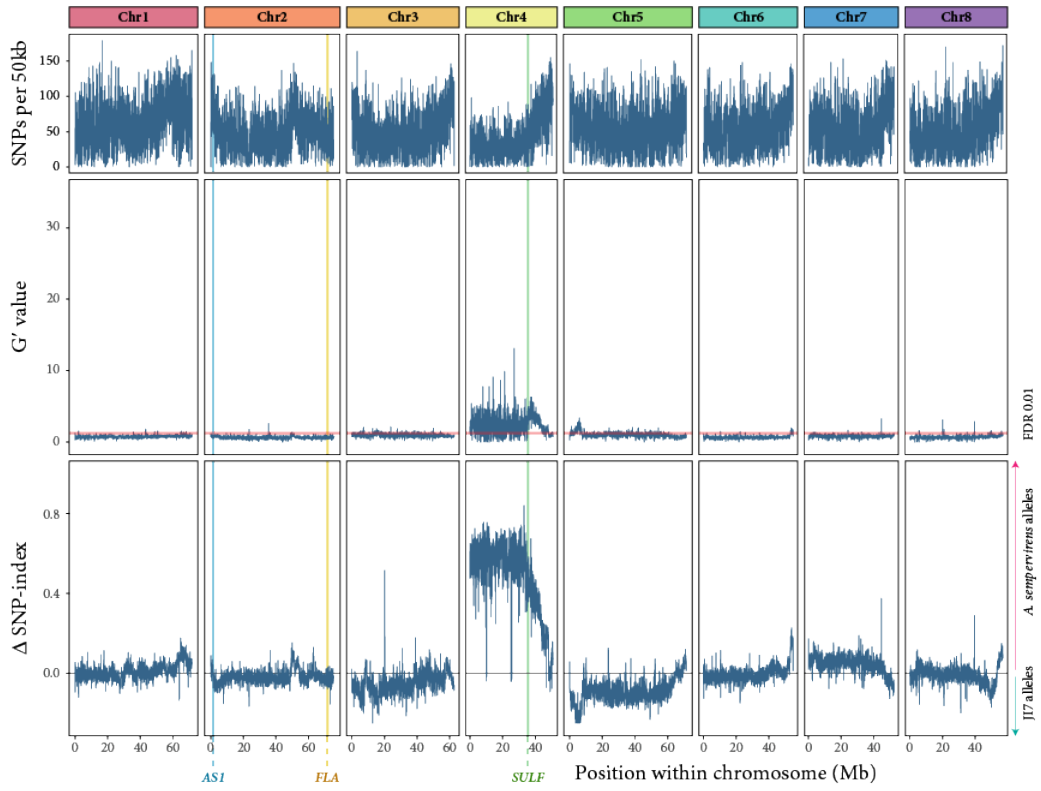


Figure 4.37 Bulked segregant analysis Manhattan plots for family H115 comparing plants that had restricted yellow pigmentation with those that had spread yellow colour. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools (Δ SNP-index). All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA* and *SULF* are indicated with vertical lines and labelled below the *x* axis.

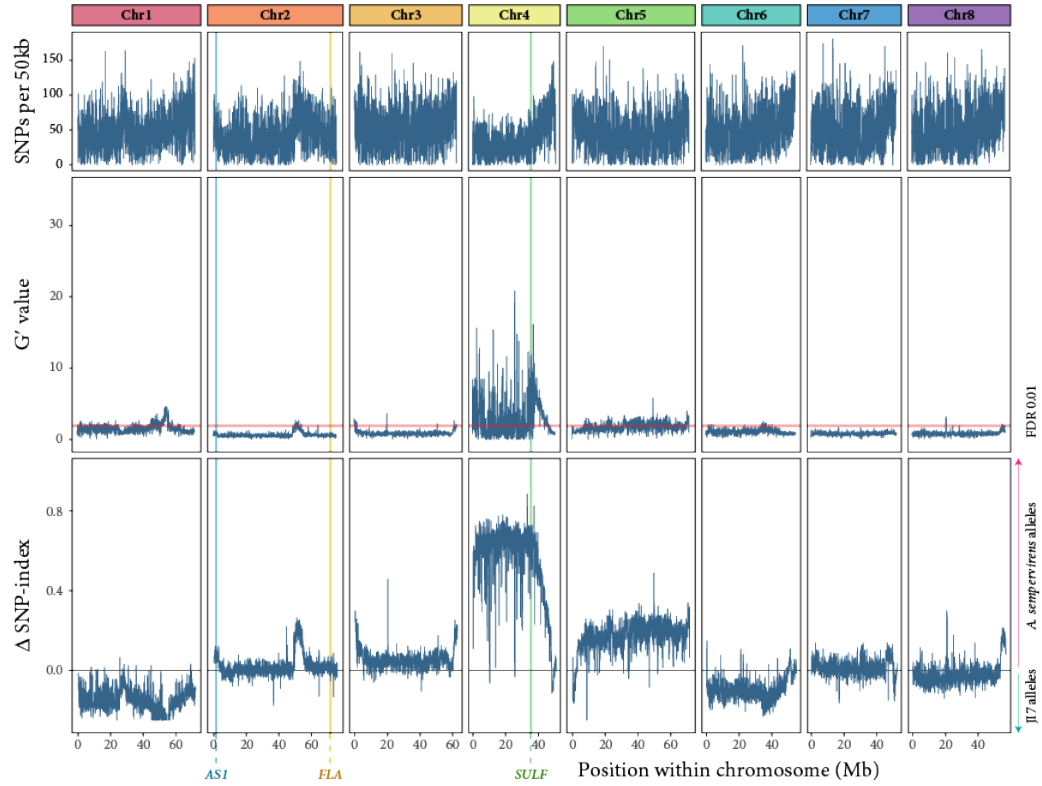


Figure 4.38 Bulked segregant analysis Manhattan plots for family H118 comparing plants that had restricted yellow pigmentation with those that had spread yellow colour. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools (Δ SNP-index). All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA* and *SULF* are indicated with vertical lines and labelled below the *x* axis.

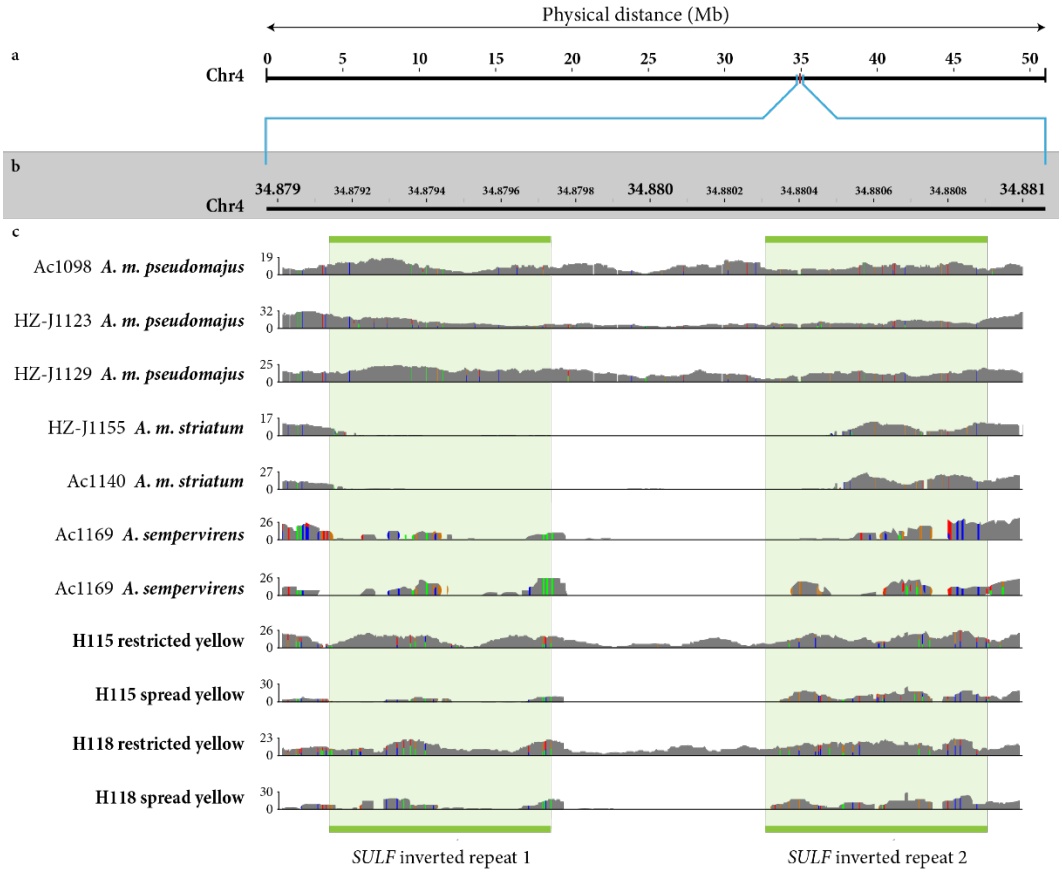


Figure 4.39 Whole genome sequencing depth of coverage along a section of chromosome 4 containing the two *SULF* inverted repeat sequences for three *A. m. pseudomajus* (restricted yellow), two *A. m. striatum* (spread yellow) individuals, two *A. sempervirens* (little or no yellow) individuals, and pooled plants with restricted yellow and spread yellow on their flowers. Panel **a** shows the whole length of chromosome 4, with the region expanded in **b** and **c** highlighted. In **c**, the sequenced individuals are shown on the left, and their coverage profiles are illustrated using panels from the Integrative Genomics Viewer (Robinson *et al* 2011), with the *y* axes showing the number of reads mapped at each position in the genome. Coloured vertical lines show positions that contain SNPs relative to the reference genome. The locations of the two *SULF* inverted repeats are shown by the green boxes.

I also looked at the *SULF* sequence of plants in H115 and H118 and compared them to those of *A. m. striatum*, *A. m. pseudomajus* and *A. sempervirens*. This showed that plants with spread yellow flowers in H115 and H118 had a deletion between the two *SULF* inverted repeats and that *A. sempervirens* had the same deletion. This deletion was smaller than that of *A. m. striatum* and did not include the inverted repeats themselves. However, the phenotype suggests that this deletion is enough to allow yellow pigmentation to spread throughout the petal lobes.

In the bulked segregant analysis of H246, comparing plants with and without yellow flower tubes, the largest peak in G' value and Δ SNP-index was once again on chromosome 2 (**Figure 4.40**). Few other genomic positions have G' values that cross the false discovery rate threshold, and none have a similar peak height to chromosome 2 positions, apart from the narrow peaks on chromosomes 4 and 5, which, as described earlier in this section, are likely assembled incorrectly.

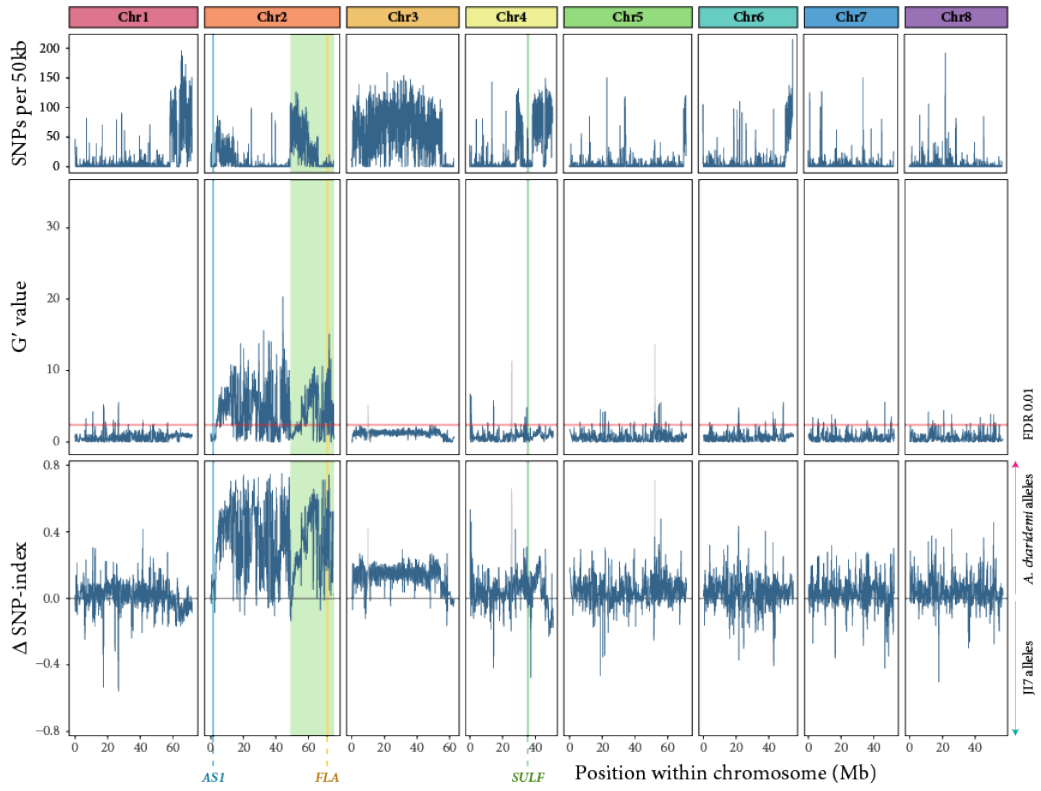


Figure 4.40 Bulked segregant analysis Manhattan plots for family H246 segregating for the presence and absence of yellow pigmentation in the tubes of the flowers. The top plot shows the number of SNPs in each window, the middle plot shows the G' value and the bottom plot shows the difference in allele frequency between the pools. All values are averaged across 50 kb windows; the lower two plots are smoothed using a tri-cube kernel function. The red line in the middle plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA* and *SULF* are indicated with vertical lines and labelled below the *x* axis.

One difference between the plot for H246 and those for H115 and H118 is that the H246 G' values are generally lower outside the peaks than the corresponding values in the other families. The SNP density across much of the chromosome is also lower, and the Δ SNP-index values are closer to 0 on most chromosomes. This may reflect

the fact that H246 is far more introgressed than H115 and H118 or, indeed, any other family used in this thesis. Through multiple generations of backcrossing, more loci have become fixed in H246 than in the other populations I have used.

As in previous results for other phenotypes in this chapter, the signal on chromosome 2 associated with the yellow phenotype excludes the *AS1* locus but includes the *FLA* locus. However, it also includes most positions on the chromosome, which leaves the result inconclusive as to whether *FLA* or another locus is responsible for the yellow tube phenotype. One hypothesis to explain this is that a *cis*-regulatory region change at *FLA* in *A. charidemi* results in yellow aurones being produced and accumulated in regions of the flower that are not yellow in other species. The alternative is that another gene, also located on chromosome 2, is responsible for yellow colour production in the flower tube of *A. charidemi*. Such a locus may encode a transcription factor that activates yellow production in the flower tube in *A. charidemi* or deactivates the pathway in other species. Alternatively, it may be transcribed as regulatory sRNAs that inhibit the activity of *FLA* or *AS1* in the flower tube in species other than *A. charidemi*. A way to test this would be to grow a larger population segregating for the same phenotype and to genotype these plants for *FLA*. If the yellow tube phenotype is regulated at the *FLA* locus itself, every plant with a yellow tube should have the $FLA^{charidemi}/FLA^{charidemi}$ genotype. If the phenotype is regulated by a locus located elsewhere on chromosome 2, if the family is large enough to ensure recombination between *FLA* and the causal locus, some plants with a yellow tube will have a $FLA7/FLA^{charidemi}$ genotype.

I also replotted the figures in this section using larger 1 Mb windows (**Figure 4.41**) as I did for J152 and J154. The largest peaks for H115 and H118 when comparing plants that have the no yellow phenotype with those that have restricted yellow, and when comparing H246 plants with and without yellow tubes, are still on chromosome 2. These peaks still exclude *AS1* and, while the signal for *FLA* is higher than the false discovery rate threshold for each comparison, there are considerably higher peaks elsewhere on the chromosome. The largest peaks for H115 and H118 when comparing plants that have the spread yellow phenotype with those that have restricted yellow is on chromosome 4, near the *SULF* locus.

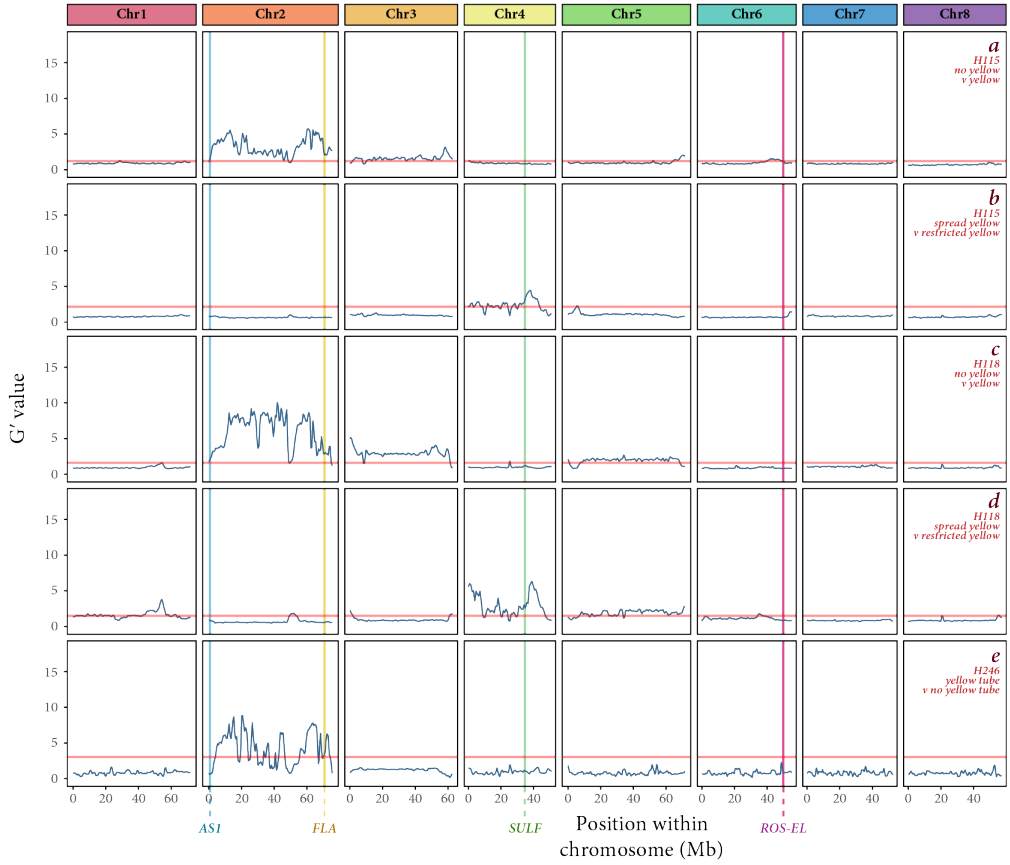


Figure 4.41 Bulked segregant analysis Manhattan plots showing G' values for H115 comparing plants that have no yellow pigmentation on the flower face with those that have restricted yellow (**a**), H115 comparing plants that have yellow pigmentation spread across the petal lobes with those that have restricted yellow (**b**), H118 comparing plants that have no yellow pigmentation on the flower face with those that have restricted yellow (**c**), H118 comparing plants that have yellow pigmentation spread across the petal lobes with those that have restricted yellow (**d**), and H246 comparing plants that have yellow flower tubes with those that do not (**e**). Note that the y axis differs from that used in previous plots. All values are averaged across 1 Mb windows and smoothed using a tri-cube kernel function. The red line in each plot represents a G' threshold corresponding to a false discovery rate of 0.01. The positions of *AS1*, *FLA* and *SULF* are indicated with vertical lines and labelled below the x axis.

4.6 Discussion

4.6.1 Origin of additional phenotypes

My initial hypothesis for work in this chapter was that yellow flower colour variation

in the *Antirrhinum* genus would have additional regulators to those already described. I looked at three different yellow phenotypes that had not been genetically characterised: yellow arc (from the *A. m. striatum* and *A. m. pseudomajus* hybrid zone), yellow tube (from *A. charidemi*) and lack of yellow on the face (from *A. sempervirens*). I also looked at a spread yellow phenotype seen when *A. sempervirens*, which has little or no yellow colour on its flowers, was crossed to a lab cultivar with restricted yellow.

All my bulked segregant analyses for this chapter showed a peak on chromosome 2, and all of these peaks included the *FLA* locus, which encodes one of the two aurone-specific enzymes. It is unclear from these results whether the genetic changes underlying yellow colour variation are in the coding region or *cis*-regulatory region of *FLA*, or whether a separate regulator is encoded at a linked locus. However, these results suggest that *FLA*, or another linked gene on chromosome 2, is responsible for much of the variation in yellow colour between *Antirrhinum* species.

My results also revealed a third allele present in the *A. m. striatum* and *A. m. pseudomajus* hybrid zone at the *SULF* locus. This locus in *A. m. pseudomajus* contains *FLA*-derived inverted repeat sequences that restrict aurone production in specific petal regions by restricting *FLA* expression (Bradley *et al* 2017). A deletion at this locus in *A. m. striatum* relative to *A. m. pseudomajus* allows yellow pigmentation to extend throughout the petal lobes. The additional allele I identified in this work appears to restrict yellow pigmentation conferred by the *A. majus* JI7 allele of *FLA* to the arc region of the dorsal petals and the foci region of the ventral petals. This suggests a weaker restriction of *FLA* than the *A. m. pseudomajus* allele, which restricts yellow pigmentation to just the foci. The frequency of this 'yellow arc' *SULF* allele among plants in the hybrid zone region is unknown, but future experiments looking at this could take advantage of hybrid zone tissue already collected. KASP primers could be developed to discriminate between the *A. m. pseudomajus* and 'yellow arc' *SULF* allele in individuals from this population.

One technique that could help to determine the regulatory basis of yellow colour variation more accurately is RNA sequencing. Sequencing RNA from different petal tissues in plants with different yellow colour phenotypes would reveal which genes are expressed in which tissues. Dissecting the tube and lobe regions of plants with and without strong yellow pigmentation would allow comparisons of gene expression between different tissues. If *FLA* is responsible for the yellow tube phenotype, for example, I would expect its expression to be higher in the tube region of plants

showing this phenotype compared with the same region in plants without the phenotype and compared with non-yellow regions of the same flowers. RNA sequencing would also reveal whether any additional, unknown genes are differentially expressed between tissues of interest. This makes it a more robust method of studying gene expression than targeted analyses such as qPCR, where primers would need to be developed for each gene of interest.

4.6.2 The role of biosynthetic genes in natural variation

One key difference between magenta and yellow colours in *Antirrhinum* is the way in which they are regulated. Magenta colour is regulated by transcription factors, which activate and suppress anthocyanin structural genes in different tissues (Rausher 2006, Schwinn *et al* 2006, Tavares *et al* in review). Regulation of yellow colour, however, appears to be linked to the regulatory sRNA locus *SULF* (Bradley *et al* 2017) and the aurone structural gene *FLA* (Boell *et al* unpublished results).

This difference in the way aurones and anthocyanins are regulated may be because of the pleiotropic nature of the anthocyanin structural genes. Anthocyanins are used by plants for defence against a host of biotic and abiotic stresses (Koes *et al* 1994). There has been less work on the biological significance of aurones beyond flower colour, but their relative rarity among plants and their absence in all clades except the flowering plants (Rausher 2006) suggests that they are less important for defence than anthocyanins. Aurones are also believed to have evolved more recently than anthocyanins. The anthocyanin pathway and its regulation have evolved piecemeal over millions of years (Rausher 2006), with at least eight enzymes required for cyanidin 3-rutinoside biosynthesis in *Antirrhinum* (Martin *et al* 1991). The shorter pathway and shorter evolutionary timescale of aurone biosynthesis may mean that yellow colour by aurones requires a simpler regulatory mechanism.

4.6.3 Lack of recombination on chromosome 2

A common feature seen in all the bulked segregant analyses for this chapter is a lack of recombination along a large interval on chromosome 2. Peaks on this chromosome were consistently nearly chromosome-wide, suggesting that recombination is suppressed across a 55-65Mb interval. This suppressed recombination could be a sign of a chromosomal inversion between JI7 and the accessions used in crosses for this chapter. Inversions, formed when an interval breaks apart from the rest of the

chromosome and is reinserted in the reverse orientation, lead to suppressed recombination in heterozygotes because gametes are not balanced (Kirkpatrick and Barton 2006, Kirkpatrick 2010). Inversions sometimes contain adaptive combinations of genes and the resulting suppressed recombination is believed to protect these combinations (Twyford and Friedman 2015). Determining the cause of the reduced recombination on chromosome 2 in *Antirrhinum* will first require better characterisation of recombination rates across the chromosome using genetic markers located at regular intervals.

4.6.4 Contribution to the understanding of the *Antirrhinum* genome

This work used a new version of the *Antirrhinum* genome assembly (known informally as *A. majus* GenomeV2, Xue *et al*, unpublished). This version of the genome comprises eight scaffolds, each corresponding to one of the eight chromosomes of *Antirrhinum*. However, during my work, I encountered several genomic regions on many different scaffolds where allele frequency differences identified in the BSA were in strong contrast to surrounding regions. For example, several narrow 'peaks' in allele frequency difference are seen on chromosomes 4, 5 and 6, while a section of chromosome 2 shows a signal 'dip' around 50Mb. I concluded that the anomalous narrow peaks were likely misplaced in the genome assembly and instead belong on chromosome 2 and that the dip in the chromosome 2 signal was the result of part of the scaffold being assembled in the incorrect orientation. This information will be used to inform future improvements to the genome assembly, and these changes will be incorporated into future genome releases.

5 Discussion

5.1 Summary of the work presented in this thesis

In this thesis, I have used a combination of bulked segregant analysis, fine mapping of traits and whole genome sequencing of natural accessions to look at the genetic basis and evolution of flower colour in *Antirrhinum*. I tested whether each flower colour trait (ie accumulation of anthocyanins and of aurones) is centrally controlled by a single locus in different species, or is a trait with dispersed genetic control with many loci contributing towards a phenotype. Six flower colour phenotypes were mapped to three loci. I concluded that there are fewer loci regulating flower colour than there are colour phenotypes seen, but that there are also several distinct loci involved in controlling each trait. This means that neither hypothesis is fully correct and that genetic control of flower colour in *Antirrhinum* lies somewhere between being centralised and dispersed.

5.2 Flower colour in *Antirrhinum* is regulated by fewer loci than the number of different phenotypes seen

Across *Antirrhinum*, there are many distinct flower colour and pattern phenotypes.

Species can broadly be categorised as having magenta, white or yellow flowers. But within these categories, more variation exists. Magenta flowers can have a white ventral patch near the flower opening ('white face') or the amount of anthocyanin that accumulates can be reduced, resulting in a paler pink colour. Most white flowers have strong yellow foci that are thought to guide pollinators to the precise part of the flower where the upper and lower lobes meet. In white flowers, there are also usually magenta-coloured venation patterns in their dorsal lobes, but different species vary for the degree to which these patterns are restricted. Yellow flowers often have these vein patterns too. And some species, notably *A. charidemi*, have yellow colouration in the tube of the flower, while most species have unpigmented or magenta tubes.

Some of these phenotypes had already been genetically characterised. Production of magenta anthocyanins in the petal epidermis had been mapped to the *ROSEA* (*ROS*) locus and production of the same pigments in tissues overlying dorsal veins to the *VENOSA* (*VE*) locus (Schwinn *et al* 2006, Shang *et al* 2011). Later, the restriction of the *ROS* and *VE* phenotypes to the central region of the flower was mapped to the *ELUTA* (*EL*) locus, which is linked to the *ROS* locus (Tavares 2014, Tavares *et al* in review). Likewise, the restriction of the yellow aurone pigment to the flower foci was mapped to the *SULFUREA* (*SULF*) locus in *A. majus* and *A. m. pseudomajus* (Bradley *et al* 2017) and to the *FLAVIA* (*FLA*) locus in *A. molle* (Boell and Bradley, unpublished results). But the additional variation seen in other species – lack of yellow in *A. sempervirens*, yellow tube in *A. charidemi*, white face (localised lack of magenta) in *A. m. pseudomajus* – remained unexplained. More flower colour variation, hidden because of epistasis, was also seen when species were crossed to lab cultivars to characterise their phenotypes and genotypes, including the yellow arc phenotype from the *A. m. striatum* and *A. m. pseudomajus* hybrid zone, and another white face (localised lack of magenta) phenotype from *A. molle*.

Bulked segregant analysis of the white face phenotype from *A. m. pseudomajus* revealed that an allele from the subspecies at the *ROS-EL* locus is responsible for the phenotype. This was confirmed by individual genotyping, which showed that the *A. m. pseudomajus ROS* allele (ROS^P) appears to be more tightly linked to the phenotype than EL^P, although the recombination rate I calculated cast some doubt over the accuracy of these markers. If this result is correct, it means that ROS^P encodes a ROS transcription factor that is expressed differently from that of the *A. majus* lab cultivar JI7 and leaves part of the face of the flower without magenta pigmentation. If *EL* is the causal locus, the EL transcription factor is produced in *A. m. pseudomajus* and EL^P

is not non-functional as previously thought, although it has a very different expression pattern from the *EL*^{S} allele in *A. m. striatum*. Whichever is correct, there are at least three *ROS* or *EL* alleles with different effects possible in the genus (**Figure 5.1**).

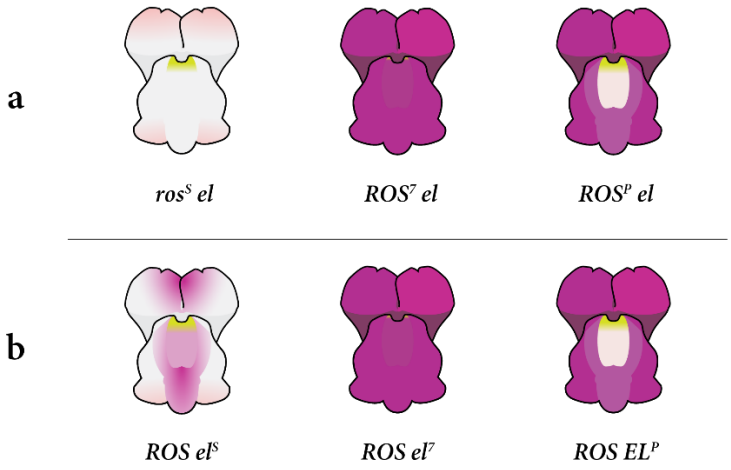


Figure 5.1 Explaining the white face phenotype if *ROS* (**a**) or *EL* (**b**) is the causal locus associated with the phenotype. A superscript S refers to an allele from *A. m. striatum*, 7 refers to an allele from the *A. majus* variety JI7 and P refers to an allele from *A. m. pseudomajus*. All genotypes depicted are homozygous.

Bulked segregant analysis of the white face phenotype from *A. molle*, however, showed that white face variation in this species is not regulated by *ROS* or *EL* – in this case, the phenotype associated with an *A. molle* allele on chromosome 2, with a possible modifier on chromosome 5. One gene located on this chromosome is *CHALCONE ISOMERASE* (*CHI*), which encodes the chalcone isomerase enzyme involved in the biosynthesis of anthocyanins and other flavonoids. It may be possible that mutations in the *cis*-regulatory region of this gene cause differences in where this gene is expressed, although transcription factors interacting with the anthocyanin pathway have previously only been shown to interact with later stages (**Figure 5.2**) (Schwinn *et al* 2006). Neither does the causal locus of the white face phenotype in *A. molle* appear to be a homologue of the *LAR1* gene, which produces a similar phenotype in *Mimulus lewisii* (Yuan *et al* 2016). The *Antirrhinum* homologue of this gene is located on chromosome 1, where no peak is seen. Whatever the genetic basis of this phenotype in *A. molle*, however, it is different from that seen in *A. m. pseudomajus*, which suggests that there may be more than one way of producing a white face in *Antirrhinum* flowers.

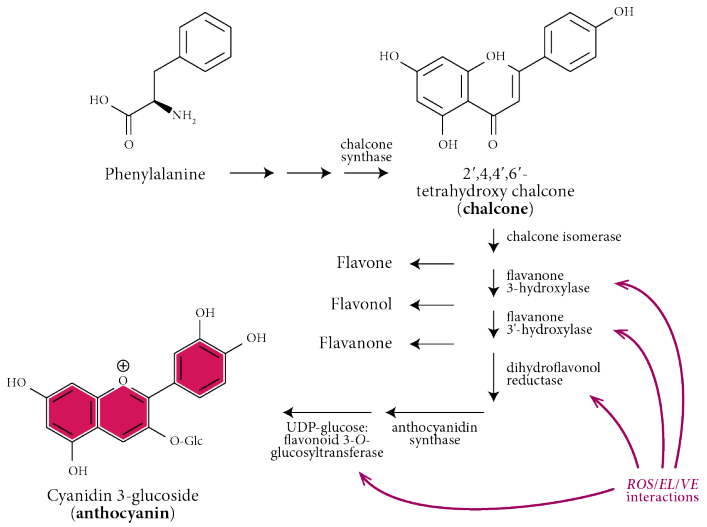


Figure 5.2 The biosynthetic pathway of cyanidin-3-glucoside, the anthocyanin produced in *Antirrhinum* flowers, annotated with the parts of the pathway with which the ROS, EL and VE transcription factors interact. Drawn using information from Schwinn *et al* (2006).

Three yellow colour phenotypes from three different species mapped to chromosome 2: increased yellow production (as seen in the arc region) in an F2 between JI7 and an accession from the *A. m. pseudomajus* and *A. m. striatum* hybrid zone; the lack (or near lack) of yellow pigmentation in the flowers on *A. sempervirens*; and the yellow pigmentation seen in the flower tubes of *A. charidemi*. Two explanations can be given for this. Firstly, the same gene – likely the chalcone glucosyltransferase-encoding biosynthetic gene *FLAVIA* (*FLA*) (Boell *et al*, unpublished results) – is responsible for several yellow phenotypes. Different species may vary at the *FLA* promoter region, leading to different *FLA* expression patterns. Secondly, chromosome 2 may contain several genes involved in the regulation of yellow flower colour, making this a 'flower colour chromosome'.

Finally, two phenotypes were associated with the *SULFUREA* (*SULF*) locus on chromosome 4. In plants segregating for a yellow arc phenotype, seen in an F2 from a cross between JI7 and an accession from the *A. m. striatum* and *A. m. pseudomajus* hybrid zone, there was no segregation at the *SULF* locus, but rather the phenotypic variation mapped to chromosome 2. However, the increase in yellow pigment production caused by differences on chromosome 2 produced a pattern that had not been observed in the wild. Inspection of the *SULF* sequence revealed a deletion within this locus in both pools relative to the reference genome. *A. m. striatum* also has a deletion at *SULF*, leading to bright aurone production throughout the corolla lobes,

whereas *A. m. pseudomajus* does not, leading to restriction of yellow pigment production to the flower foci. But the yellow arc *SULF* deletion does not match that of *A. m. striatum*, suggesting that a third *SULF* allele can be found in the hybrid zone.

The other phenotype that mapped to the *SULF* locus was the spread yellow colour seen segregating in an F2 from a cross between JI7 and *A. sempervirens*. *A. sempervirens* flowers are white, with little or no yellow pigmentation. However, my analyses showed that these plants with spread yellow in this population had a deletion relative to the reference genome between the two *SULF* inverted repeats – a deletion shared by the *A. sempervirens* accession used in the cross.

5.3 Magenta flower colour is regulated by different alleles of genes encoding transcription factors

My results showing that either *ROS* or *EL* is responsible for the white face phenotype suggest that at least three alleles can exist at the causal locus. If *EL* is the causal locus, there is one JI7 allele that leads to full magenta pigmentation, a second *A. m. striatum* allele that either leads to restriction of magenta to the centre of the flower, and a third *A. m. pseudomajus* allele that produces a white face phenotype. If *ROS* is responsible, the same explanation stands, except that the *A. m. striatum* allele eliminates magenta pigmentation from the petal lobes, except for at the independently-regulated veins.

There are more experiments that can be carried out to test whether *ROS* or *EL* is responsible for the white face phenotype. Firstly, fine mapping using more genetic markers and larger plant populations would allow recombination events closer to the locus to be characterised and the phenotypes of recombinants determined. Using the *A. m. striatum* and *A. m. pseudomajus* hybrid zone as a natural laboratory, and markers dispersed along an interval containing *ROS-EL*, Tavares *et al* (in review) were able to map *EL* based on the phenotype of *A. m. striatum* to a ~50 kb interval. However, if this technique were to be used to finely map the white face phenotype, a natural population showing variation in the white face phenotype would be required. All *A. m. pseudomajus* accessions studied so far have shown some degree of localised lack of magenta pigmentation at the centre of the flower face. Instead, generating a larger segregating population, again using an F2 between *A. m. pseudomajus* and JI7, may be more suitable. The markers used would also need to cover the entirety of the peak seen in the bulked segregant analysis of the white face segregating population to determine whether another locus further up or downstream of *ROS* and *EL* is causal to the

phenotype.

Secondly, RNA sequencing from different parts of flowers with and without the white face phenotype will determine where in these flowers *ROS* and *EL* are transcribed. Work by Tavares *et al* (in review) showed that, in whole flower buds, EL^{S} (*A. m. striatum*) expression (**Figure 5.3 B tracks 11, 12, 13**) was significantly higher than that of EL^{7} (JI7) (**Figure 5.3 B tracks 8, 10, 14**), but there was no significant difference between EL^{7} (**Figure 5.3 B tracks 8, 10, 14**) and EL^{P} (*A. m. pseudomajus*) (**Figure 5.3 B track 9**) expression. Likewise, ros^{dorsea} (ros^{dor}) (**Figure 5.3 A track 7**) and ros^{S} (**Figure 5.3 A track 6**) expression levels were significantly lower than those of ROS^{7} (**Figure 5.3 A tracks 1 and 4**) and ROS^{P} (**Figure 5.3 A tracks 2 and 5**), but there was no overall significant difference between ROS^{P} (**Figure 5.3 A tracks 2 and 5**) and ROS^{7} (**Figure 5.3 A tracks 1 and 4**) expression. However, there was also no significant difference in expression between these two alleles and '*ROS**' – a recombinant allele containing the promoter and coding region of *ROS1* from *A. m. striatum*, but the rest of the *ROS-EL* locus from *A. m. pseudomajus*. This lack of difference in expression was despite the magenta colour produced being paler. This suggests that *ROS2* and/or *ROS3*, tandem downstream duplications of *ROS1*, may also contribute to the colour phenotype seen (Tavares *et al* in review). In $rosea^{colorata}$ mutants where *ROS2* is expressed but *ROS1* is not, anthocyanin production is reduced (Schwinn *et al* 2006), suggesting that *ROS2* is functional. It has not been shown whether *ROS3* is functional, which may mean it is a pseudogene (Tavares 2014).

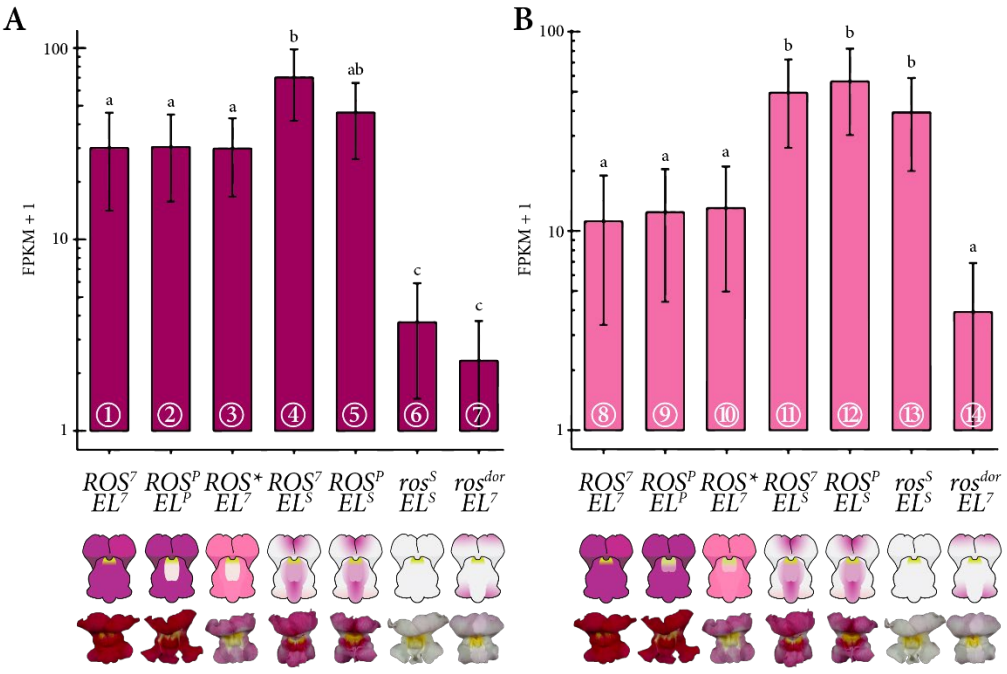
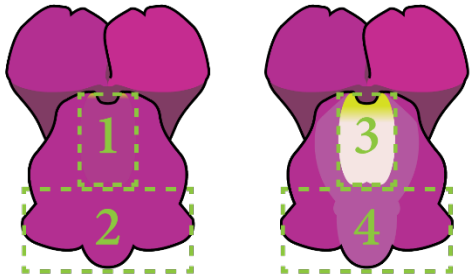


Figure 5.3 Expression estimates of *ROS1* (**A**) and *EL* (**B**) in the buds of whole flowers with different *ROS-EL* genotypes (all homozygous). The bars show the mean and 95% confidence interval of expression, measured as fragments per kilobase of transcript per million mapped reads (FKPM). The flower illustrations and photographs underneath each plot depict the phenotype seen for each *ROS-EL* haplotype written above the images. From left to right, the *ROS-EL* haplotypes are: *ROS* and *el* both from JI7; *ROS* and *el* both from *A. m. pseudomajus*; *ROS1* from *A. m. striatum* but *ROS2*, *ROS3* and *EL* from *A. m. pseudomajus*; *ROS* from JI7 and *EL* from *A. m. striatum*; *ROS* from *A. m. pseudomajus* and *EL* from *A. m. striatum*; *ros* and *EL* both from *A. m. striatum*; ros^{dor} mutation and *EL* from JI7. Samples with different letters above their bars are significantly different from each other ($q < 0.01$). Adapted from Tavares *et al* (in review).

The expression values measured by Tavares *et al* (in review), however, were for the whole corolla of flower buds, not for specific parts of the petals. Tissue-specific RNA sequencing can reveal different patterns of expression for different genes during development (Jiao *et al* 2017). *ROS* and *EL* expression overall may not differ between plants with and without the white face phenotype, but tissue-specific differences may exist. **Figure 5.4** shows the different predicted expression patterns of *ROS* and *EL* in white face and non-white face flowers if *ROS* or *EL* is responsible for the phenotype. If neither is responsible, another gene may show differential expression between the samples. These tissues have been prepared and sent for RNA sequencing, but the data has not yet been returned. The two hypotheses presented in **Figure 5.4** could also be

tested using quantitative polymerase chain reaction (qPCR), which would show the amount of transcript of each gene tested present in each sample (Stanton *et al* 2017). However, I opted for RNA sequencing because this would allow identification of all differentially expressed genes in the tissues of interest, not just those being tested. I also had limited time available to design and test new qPCR primers and chose to prioritise my BSA experiments instead; RNA sequencing does not require development of gene-specific primers.



Flower region	Expression if *EL* is causal	Expression if *ROS* is causal
1	High *ROS* Reduced *EL*	High *ROS* Reduced *EL*
2	High *ROS* Reduced *EL*	High *ROS* Reduced *EL*
3	Reduced *ROS* High *EL*	Reduced *ROS* Reduced *EL*
4	High *ROS* Reduced *EL*	High *ROS* Reduced *EL*

Figure 5.4 Predicted expression patterns of *ROS* and *EL* developing *Antirrhinum* flowers (as illustrated on diagrams of fully developed flowers) with different phenotypes depending on which of the two is causal to the phenotype.

5.4 Anthocyanin pigmentation appears to have additional regulators that have not yet been identified

When *A. molle* was crossed to JI7, around a quarter of individuals in the F2 generation had a white face phenotype very similar to that seen in *A. m pseudomajus*, while around one 16th of individuals had a stronger 'white band' pattern that covered the upper half of the ventral and lateral lobes and around 17% had a *ros*^{dor}-like phenotype. These ratios suggested that the phenotype was controlled by two genes, while the similarity of the phenotype to that previously seen in *A. m. pseudomajus* and the presence of a *ros*^{dor}-like phenotype suggested that *ROS* was one of these two genes. However, given that no peak was seen at the *ROS-EL* locus when comparing pools with and without a white face/white band phenotype, the white face seen in this population must be because of segregation at another locus.

Because the only peak seen when comparing pools that had full magenta pigmentation with those that had a strong centralised white face was on chromosome 2, it is likely that a regulator of magenta can be found on this chromosome. As described above,

this may be the result of genetic differences at or near the biosynthetic gene *CHI*, which encodes an enzyme that converts chalcone to naringenin, which is then processed further to produce anthocyanins. Changes in the coding region of *CHI* that affect the enzyme itself are unlikely here – the ability of the plants to produce anthocyanins is not affected. However, one mechanism by which flower colour can be regulated in different plant tissues without compromising the integrity of biosynthetic pathway is by mutations in the *cis*-regulatory regions of biosynthetic genes (Streisfeld and Rausher 2011, Sobel and Streisfeld 2013). Such mutations in regulatory elements – regions adjacent to coding sequences that encode instructions determining when and where the gene is transcribed by interacting with transcription factors – are common hallmarks of phenotypic variation between species (Stern and Orgogozo 2009). Through this mechanism, anthocyanin production could be locally downregulated in the face of the flower without causing pleiotropic effects – the pigments could still be produced outside this region (Wu *et al* 2013). Previously-described flower colour-regulating transcription factors in *A. majus* and *Mimulus lewisii*, however, have only been shown to interact with genes that encode enzymes downstream of chalcone isomerase in the pathway (Schwinn *et al* 2006, Wu *et al* 2013).

Another gene encoded within the chromosome 2 region where a peak is seen in J104 is *FLAVIA* (*FLA*), which encodes an enzyme involved in yellow aurone production in *Antirrhinum* flowers. *A. molle* has restricted yellow foci despite having a deletion at the *SULFUREA* (*SULF*) locus, which restricts yellow pigmentation in *A. majus* and *A. m. pseudomajus*. Instead, the pigment restriction in *A. molle* is thought to be a result of *cis*-regulatory mutations at the *FLA* locus (Boell *et al*, unpublished results). Given that double-pigmented orange *Antirrhinum* flowers are confined to hybrid zones through apparent selection against them by pollinators (Whibley *et al* 2006), it may be important for fitness that magenta and yellow pigments are kept separate on the flower. The magenta-regulating genes *ROS* and *EL* are in linkage disequilibrium on chromosome 6 in a hybrid zone between *A. m. striatum* and *A. m. pseudomajus* and work together to produce distinct phenotypes in the two subspecies (Tavares 2014). A magenta colour regulator linked to *FLA* might ensure that yellow and magenta pigments do not overlap in flowers with restricted yellow pigmentation regulated by *FLA*.

As with the white face phenotype from *A. m. pseudomajus*, one experiment that would provide useful evidence as to which gene or genes regulate the white face/band phenotype in *A. molle* is RNA sequencing of dissected corollas. This would show

which genes show different levels of expression in parts of the flower kept without pigmentation and those that are magenta-coloured.

5.5 Yellow flower colour appears to be regulated by different alleles of biosynthetic genes and of a locus transcribed as regulatory small RNAs

A common feature of nearly all my bulked segregant analyses of yellow colour variation from different *Antirrhinum* species was a signal on chromosome 2, where both enzymes involved in the conversion of chalcone to the yellow pigment aureusidin glucoside are encoded. While I was not able to map this variation precisely, owing to the low recombination on this chromosome between *A. majus* JI7 and the accessions tested (Boell and Bradley, unpublished results), absence of evidence to the contrary suggests that the biosynthetic gene *FLAVIA* (*FLA*) may be a regulator of yellow colour in several species. This gene encodes chalcone glucosyltransferase (Boell and Bradley, unpublished results), which is necessary for transportation of chalcone to the vacuole where the activity of the final enzyme in the aurone pathway takes place (Ono *et al* 2006).

The other locus to which variation in yellow flower colour was attributed was *SULFUREA* (*SULF*) on chromosome 4. Bradley *et al* (2017) showed that this locus contains inverted repeat sequences that are transcribed as regulatory small RNAs (sRNAs). These sRNAs inhibit the activity of *FLA* in specific parts of the corolla in some species with restricted yellow pigmentation such as *A. m. pseudomajus* and *A. majus*. The sister subspecies to *A. m. pseudomajus*, *A. m. striatum*, has a deletion covering part of the *SULF* locus and, as a result, *FLA* activity is not inhibited and the corolla lobes are bright yellow. My results showed that the wild parent of the F2s segregating for yellow arc – a hybrid between *A. m. pseudomajus* and *A. m. striatum* – carried a third allele of *SULF*, which has a shorter deletion than *A. m. striatum*. This allele does appear to inhibit *FLA* activity, but in a smaller proportion of the corolla than the *SULF* alleles previously studied in *A. m. pseudomajus* and *A. majus* JI7. The result in a background with the strongly-expressed JI7 allele of *FLA* is a yellow arc pattern in the dorsal lobes. This yellow arc *SULF* allele may be the result of recombination at the locus or it may be an allele that is found in some *A. m. pseudomajus* individuals.

I also showed that individuals with spread yellow flower colour in an F2 population from a cross between *A. sempervirens*, which has mostly white flowers, and *A. majus* JI7

also had a deletion in *SULF*. In fact, this deletion in the *A. sempervirens SULF* allele was similar in length to that seen in the yellow arc *SULF* allele. *A. sempervirens* and *A. m. pseudomajus* have adjacent population ranges (Wilson and Hudson 2011) and some hybridisation may have occurred, allowing introgression of the *A. sempervirens* allele into some of *A. m. pseudomajus* populations.

If changes at *FLA* – whether in its coding region or in an adjacent regulatory element – are responsible for differences in flower colour between species, the regulation of yellow pigmentation would appear to use a different mechanism from that of magenta colour. To date, all the interspecies variation in magenta flower colour in *Antirrhinum* has mapped to MYB-like transcription factors rather than the biosynthetic genes themselves (Schwinn *et al* 2006, Shang *et al* 2011, Tavares 2014, Tavares *et al* in review). Reviewing the literature documenting the genetic basis of flower colour transitions, Streisfeld and Rausher (2011) determined that fixed differences in flower colour intensity between species are characterised by changes at genes that encode transcription factors.

One notable difference between the anthocyanin and aurone pathways is the number of steps involved. Anthocyanins are produced using a multistep pathway that involves at least six enzymes between chalcone and the magenta pigment in its final form. With a separate gene encoding each enzyme, there are several points in the pathway with which transcription factors can interact to regulate flower colour (**Figure 5.5**). The aurone pathway is much shorter, containing just two enzymes that first glucosulate chalcone for vacuolar import and then convert the chalcone glucoside to aurone (Ono *et al* 2006) (**Figure 5.5**). Because there are fewer places in this pathway where transcription factors might act, regulation using that mechanism may not be possible.

Another important difference between aurones and anthocyanins is their distribution among plants. Plants as evolutionary ancient as bryophytes (liverworts, mosses and hornworts) produce anthocyanins, and the same biosynthetic pathway is used across land plants (Campanella *et al* 2014). Aurones, however, have a much more restricted distribution and are only found in a small number of taxa (Tanaka *et al* 2008). This suggests that aurones may have evolved more recently than anthocyanins. While plants have had tens of millions of years to evolve complex regulatory mechanisms for anthocyanins, the shorter evolutionary timescale of aurones may mean that there has not been enough time for a sophisticated system of regulation to evolve. Instead, plants whose flowers are coloured by aurones may have to rely on changes at biosynthetic

genes and loci encoding regulatory sRNAs in order to regulate where these pigments accumulate.

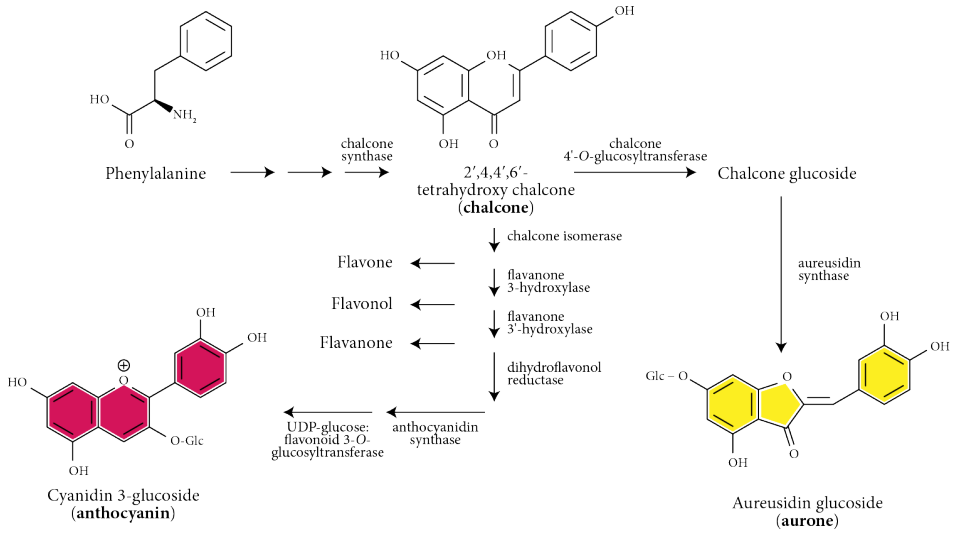


Figure 5.5 Combined biosynthesis pathways of cyanidin 3-glucoside (anthocyanin) and aureusidin glucoside (aurone). The structures of intermediate compounds between chalcone and the final pigments are not shown. Adapted from Falcone Ferreyra *et al* (2012) and Ono *et al* (2006).

5.6 Reduced recombination across chromosome 2 in *Antirrhinum*

In my analyses of flower colour regulation in different species, one common feature seen when these species were crossed to JI7 was a low rate of recombination on chromosome 2. In most of my analyses, a large chromosomal interval between the ~4Mb and ~48Mb positions, and another between the ~53Mb and ~75Mb positions, appeared to have reduced recombination. Visual analysis of signals on this chromosome further suggests that the final third of the chromosome, between ~48Mb and ~75Mb, is assembled in the wrong orientation, as the signal at the end of the chromosome is consistently level with that at the ~48Mb position. If this part of the chromosome is indeed misassembled, it would suggest a single interval of suppressed recombination 55-65Mb in length. Studies of genomic divergence between *A. m. striatum* and *A. m. pseudomajus* have also shown that recombination on this chromosome is similarly suppressed in heterozygotes between these two subspecies. Alleles on chromosome 2 in *A. m. striatum* are similar to those seen in JI7 (Boell *et al*, unpublished results). These results suggest that there are two distinct chromosome 2 haplotypes in *Antirrhinum* that do not recombine with each other.

One well-characterised source of suppressed recombination in heterozygotes between species or populations is chromosomal inversions, where part of a chromosome in one population has a different order relative to another population (Kirkpatrick and Barton 2006). These genomic features are caused when an interval between two points on a chromosome breaks apart and is reinserted in the reverse orientation. Recombination is suppressed in heterozygotes through a loss of balanced gametes (Kirkpatrick 2010). Chromosomal inversions that distinguish species or subspecies have been observed in several study systems, including *Drosophila* fruit flies (Krimbas and Powell 1992), *Anopheles* mosquitos (Coluzzi *et al* 2002) and *Heliconius* butterflies (Joron *et al* 2011). In plants, Twyford and Friedman (2015) showed that a ~6.5Mb-long chromosomal inversion underlies life history differences between annual and perennial *Mimulus guttatus* ecotypes. The *M. guttatus* inversion is believed to protect an interval containing several adaptive genes from recombination, thus providing an adaptive advantage (Twyford and Friedman 2015).

In *Antirrhinum*, a chromosomal inversion on chromosome 2 may serve a similar role in preserving linked adaptive traits to that seen in *M. guttatus*. Such an inversion would not necessarily protect loci involved in regulating flower colour, but *A. m. pseudomajus* and *A. m. striatum* are very closely related and differ for few described traits other than flower colour (Whibley *et al* 2006, Bradley *et al* 2017). Because *FLA* is located at one end of the interval with suppressed recombination and appears to be important in the determination of flower colour in *Antirrhinum*, one or more similarly important genes may be found at the other end of the interval. Although *AS1* is located at this end of the chromosome, it does not appear to be included in this low recombination interval, suggesting that there may be another flower colour gene in this region. Alternatively, there may be an entirely different reason for the low recombination on this chromosome. More work using larger populations and additional markers along chromosome 2 is needed to determine whether any recombination points can be identified within this interval and to measure recombination rates across the chromosome.

5.7 More phenotypic variation in flower colour in *Antirrhinum* remains to be explained

In this thesis, I have described and genetically mapped different magenta and yellow flower colour phenotypes seen in the *Antirrhinum* genus. However, some additional

species of *Antirrhinum* show more variation in flower colour that has not been studied. Several species, including *A. graniticum* and *A. barrelieri*, have pale pink flowers where the anthocyanin colouration is less intense than in magenta-flowered species such as *A. m. pseudomajus*. *A. graniticum* shows polymorphism in flower colour between populations – at some locations, flowers are pink, while in others they are white; some locations also show variation within populations (**Figure 5.6**).

Figure 5.6 Variation seen in the intensity of magenta pigmentation in *A. graniticum*. Some populations had pink flowers (**a**) while others had white flowers (**b**). One population was seen to show a gradient of pigment intensities between the two phenotypes (**c**). It is not known whether this difference in colour is genetic or because of environmental factors.

A similar phenotype was seen by Tavares *et al* (in review) in a hybrid plant between *A. m. striatum* and *A. m. pseudomajus* that showed recombination downstream of the *ROS1* coding sequence. This plant carried the *ROS1* allele of *A. m. striatum*, which produces no magenta in its flowers, and the downstream sequence containing *ROS2*, *ROS3* and *EL* of *A. m. pseudomajus*. This suggests that differences within the promoter or coding

region of *ROS1* can change magenta intensity without eliminating it. It would be interesting to explore whether *A. graniticum* shows differences at *ROS1* compared to *A. m. pseudomajus*, and whether there are genetic differences at *ROS1* or downstream in *A. graniticum* plants with different magenta intensities in their flowers. The intensity of flower colour can also be affected by the shape of the petal cells where the pigments accumulate. Cell shape can change the amount of light that can penetrate the petal cells, thus changing the proportion of light absorbed by pigments (Noda *et al* 1994). It is not known whether *A. graniticum* differs in the shape of its petal cells compared to *A. m. pseudomajus*.

Other species, such as *A. siculum*, show differences in yellow pigment intensity in different parts of the corolla lobes. In *A. siculum*, the face of the flower is a much brighter yellow than the rest of the petals (**Figure 5.7**). Previous variation in yellow colour has mapped either to chromosome 2 or to *SULFUREA*. Variation between *A. siculum* and other yellow-flowered species may be regulated by one of these genes or it may involve a separate locus not yet described.

Figure 5.7 *Antirrhinum siculum* flowers. Photograph by Enrico Coen.

One of the phenotypes I studied in this thesis was a yellow pigmentation in the tube of *A. charidemi* flowers. *A. charidemi* is the only species where this phenotype has been

described, although a similar phenotype has been observed in F2 populations between plants from the *A. m. striatum* and *A. m. pseudomajus* hybrid zone and JI7. However, many descriptions of flower colour in *Antirrhinum* species refer only to the petal lobes, making no reference to the colour of the flower tube (Mather 1947, Tastard *et al* 2012). Subtle differences in yellow are also difficult to score, with the observed intensity varying in different light conditions. Much phenotyping work on yellow variation included in this thesis has relied on visualising different hues separately in photographs because of this. An analysis of the flower colours of different *Antirrhinum* species using this technique may reveal addition variation in flower colour not previously described, including in tube colour.

5.8 Bulked segregant analysis is a useful tool for mapping genetic variation from wild plant populations but it has its limitations

For this thesis, I used whole genome sequencing-based bulked segregant analysis to map species-derived variation in segregating plant populations to the *Antirrhinum* reference genome. This proved to be a very useful tool that allowed me to genotype millions of sites across the genome concurrently without having to develop markers in advance. Using this technique, I was able to map six different flower colour phenotypes from four species and suggest possible mechanisms by which flower colour is regulated in those species.

However, in most of my analyses, I saw more than one peak in allele frequency difference between species, even where phenotypic segregation ratios suggested that only one was linked to the phenotypes differing between the pools. One likely reason for this is the low number of individuals used in some of the bulks, with one bulk containing as few as 13 individuals. The sizes of the bulks used in comparable studies ranges from 20 to 100 individuals (Takagi *et al* 2013, Yuan *et al* 2013, Friedman *et al* 2015, Song *et al* 2017), and it is generally acknowledged that increasing the size of the bulks minimises the variation that affects allele frequency difference estimates (Magwene *et al* 2011).

Another issue with some of my bulked segregant analyses was that some individuals had been placed into the wrong bulk. In my analyses of L124 in chapter 3, for example, one sample was found to contain several individuals that had been misgenotyped for the *ROS-EL* locus. Because of the nature of bulked segregant analysis pooling, where several individuals are combined according to phenotype, the ability to discriminate

between individuals in the sequenced data is lost. One alternative to bulked segregant analysis that avoids this problem is multiplexed shotgun genotyping (Andolfatto *et al* 2011). This technique differs from bulked segregant analysis at the preparation of sequencing libraries stage. Instead of preparing one DNA library for all individuals together, a library is prepared for each individual separately and a unique barcoded adapter is used for each sample. This allows bioinformatic identification of individuals after sequencing (Andolfatto *et al* 2011).

The method I used to get around the problem of not being able to identify individuals in my bulked segregant analysis results was to follow these analyses with genotyping of individual plants, each of which was scored for the flower colour trait segregating in that population. This meant that bulked segregant analysis was used to identify regions of interest and fine mapping techniques – KASP genotyping in this case – used to look closer at each of these regions and resolve issues at ambiguous loci. Unfortunately, I faced several issues with this technique too, such as several of the markers designed not being polymorphic between samples with different flower colour. I also saw overrepresentation of heterozygotes in some of my samples, which likely led to an overestimation of the recombination rates between the markers used at *ROS* and *EL*. This could be avoided in the future by using a greater number of markers at each locus and confirming results using an alternative genotyping method such as amplicon sequencing.

5.9 Future experiments that could expand on these results

One surprising result shown in this work is the apparently different genetic basis of the two phenotypically similar white face traits, one from *A. m. pseudomajus* and the other from (but not seen in) *A. molle*. While the white face phenotype of *A. m. pseudomajus* mapped to the *ROS-EL* locus on chromosome 6, the phenotype seen in an F2 from a cross between *A. molle* and *A. majus* mapped to chromosome 2. Thus, it appears that magenta can be regulated in two ways that produce similar results.

But *A. m. pseudomajus* is far from the only *Antirrhinum* species where a white face phenotype can be found. As discussed in chapter 3, most *Antirrhinum* species with magenta flowers show localised lack of magenta colour around the face region. An important future experiment will be to determine which – if any – of the two loci described in this thesis underlie the white face phenotype seen in these other species. This could be tested by crossing these additional species to *A. majus* to generate more

F2 populations segregating for the phenotype, and then to perform bulked segregant analysis using phenotypic extremes from this F2. Another way would be to genotype the F2s using markers developed at the *ROS-EL* locus at sites on chromosome 2. These markers could be developed using the whole genome sequencing data already available for many of these species and looking for SNPs that discriminate between *A. majus* and the species being analysed.

Another unexpected result was the identification of a previously unseen *SULF* allele in the family segregating for the yellow arc phenotype. The male parent of this family came from the hybrid zone between *A. m. striatum* and *A. m. pseudomajus*, suggesting that the allele came from one of these subspecies. However, *SULF* had already been characterised in both *A. m. striatum* and *A. m. pseudomajus*, with the *A. m. pseudomajus* allele restricting yellow pigmentation to the flower foci and the *A. m. striatum* allele being non-functional, allowing yellow pigmentation to spread. The discovery of the phenotypically intermediate yellow arc *SULF* allele suggests that more variation exists in these subspecies than has previously been described. An interesting next step will be to determine how prevalent this yellow arc *SULF* allele is in and around the hybrid zone. Leaves have been collected from thousands of hybrid zone plants every year since 2009 as part of a separate project. This means it would be possible to sample plants from many different geographic locations near the hybrid zone and determine their *SULF* haplotypes without a new collection effort.

One consistent issue in all my bulked segregant analyses was that all analyses showing a peak on chromosome 2 also showed narrow peaks on chromosomes 3, 4 and 5. The loci contained on these three peaks appear to be genetically linked to chromosome 2 and are therefore likely to be misassembled. However, their exact correct positions on chromosome 2 are unknown, and more work on this assembly will be required to determine this. The misassembly of these regions also raises questions about the rest of the genome, and there may be other regions that are misassembled.

6 Conclusions

In this thesis, I tested two alternate hypotheses to explain the genetic basis of variation in flower colour in *Antirrhinum*:

1 Each flower colour trait is regulated centrally at one locus. Different alleles at these loci result in different phenotypes. If this hypothesis is correct, variation in each trait will map to one locus.
2 Flower colour regulation is dispersed across many unlinked loci. All differences in flower colour are regulated independently. If this hypothesis is correct, variation in each trait will map to a different locus.

My results fit neither of these hypotheses perfectly. The *ROS-EL* locus appears to be the major regulator of magenta pigmentation in *A. m. striatum*, *A. m. pseudomajus* and the *A. majus* lab cultivar JI7. Activation of magenta pigmentation (Schwinn *et al* 2006), restriction of magenta pigmentation to centre of the flower (Tavares *et al* in review) and suppression of magenta pigmentation in the face region (this thesis) all map to the same locus on chromosome 6. However, variation in magenta pigmentation in other *Antirrhinum* species may be regulated separately – the white face phenotype seen in an F2 population from a cross between *A. molle* and JI7 did not map to chromosome 6. Likewise, variation in yellow pigmentation from several *Antirrhinum* species mapped

to chromosome 2, which appears to behave as one locus with little recombination happening on this chromosome in crosses between species or subspecies. However, the unlinked *SULF* locus also regulates magenta variation between the same species.

My results suggest a form of flower colour regulation intermediate to the two hypotheses I proposed is seen in the *Antirrhinum* species I studied. Restriction/spread of magenta appears to be regulated at one locus in *A. majus* and its subspecies, but other *Antirrhinum* species may have different mechanisms for regulating anthocyanin pigmentation. Restriction/spread of yellow is regulated at two loci: *SULF* and chromosome 2.

Yellow and magenta pigmentation are also regulated in very different ways from each other. Anthocyanin pigmentation is regulated by transcription factors, which activate and suppress the anthocyanin structural genes in different parts of the corolla. Yellow pigmentation, however, does not appear to be regulated by transcription factors. Instead, yellow variation is regulated at aurone structural gene loci (either in coding regions or in *cis*-regulatory regions) and at loci that are transcribed as regulatory small RNAs. This may be because the anthocyanin structural genes are pleiotropic – mutations in these genes could affect anthocyanin production throughout the plant, which may impact plant defence and other biological functions. This difference in regulation may also be because evolution of aurone biosynthesis appears to be more recent than that of anthocyanins – aurones are only found in some flowering plant taxa, whereas anthocyanins are found across most land plants.

7 References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403-410. doi:10.1016/S0022-2836(05)80360-2.

Altshuler D, Daly MJ & Lander ES (2008) Genetic mapping in human disease. *Science*, **322**(5903): 881-888. doi:10.1126/science.1156409.

Ananga A, Georgiev V, Ochieng J, Phills B & Tsolova V (2013) Production of anthocyanins in grape cell cultures: a potential source of raw material for pharmaceutical, food, and cosmetic industries. *The Mediterranean Genetic Code - Grapevine and Olive*, Poljuha D and Sladonja B (Eds), InTechOpen, London, UK.

Anderson EC, Skaug HJ & Barshis DJ (2014) Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, **23**(3): 502-512. doi:10.1111/mec.12609.

Anderson JT, Willis JH & Mitchell-Olds T (2011) Evolutionary genetics of plant adaptation. *Trends in Genetics*, **27**(7): 258-266. doi:10.1016/j.tig.2011.04.001.

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T & Stern DL (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping.

Genome Research, **21**(4): 610-617. doi:10.1101/gr.115402.110.

Aronesty E (2011) ea-utils : command-line tools for processing biological sequencing data. Expression Analysis, Durham, NC, US. Available at https://github.com/ExpressionAnalysis/ea-utils.

Barton NH & Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**(6242): 497-503. doi:10.1038/341497a0.

Baxter SW, Papa R, Chamberlain N, Humphray SJ, Joron M, Morrison C, ffrench-Constant RH, McMillan WO & Jiggins CD (2008) Convergent evolution in the genetic basis of müllerian mimicry in *Heliconius* butterflies. *Genetics*, **180**(3): 1567-1577. doi:10.1534/genetics.107.082982.

Boell L, Bradley D, Copsey L, Couchman M, Tavares H, Whibley A & Coen E (unpublished results). John Innes Centre.

Bradley D, Xu P, Mohorianu I-I, Whibley A, Field D, Tavares H, Couchman M, Copsey L, Carpenter R, Li M, Li Q, Xue Y, Dalmay T & Coen E (2017) Evolution of flower color pattern through selection on regulatory small RNAs. *Science*, **358**(6365): 925-928. doi:10.1126/science.aao3526.

Bradshaw HD & Schemske DW (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, **426**: 176. doi:10.1038/nature02106.

Broad Institute (2018a) The Genome Analysis Toolkit. The Broad Institute of MIT and Harvard, Cambridge, MS, US. Available at https://software.broadinstitute.org/gatk/.

Broad Institute (2018b) Picard Tools. The Broad Institute of MIT and Harvard, Cambridge, MS, US. Available at http://broadinstitute.github.io/picard/.

Buer CS, Imin N & Djordjevic MA (2010) Flavonoids: new roles for old molecules. *Journal of Integrative Plant Biology*, **52**(1): 98-111. doi:10.1111/j.1744-7909.2010.00905.x.

Campanella JJ, Smalley JV & Dempsey ME (2014) A phylogenetic examination of the primary anthocyanin production pathway of the Plantae. *Botanical Studies*, **55**(1): 10. doi:10.1186/1999-3110-55-10.

Charlesworth B & Charlesworth D (2010) *Elements of Evolutionary Genetics*. Roberts and

Company Publishers, Greenwood Village, CO, US, ISBN: 978-0-9815194-2-5.

Chase MW & Reveal JL (2009) A phylogenetic classification of the land plants to accompany APG III. *Botanical Journal of the Linnean Society*, **161**(2): 122-127. doi:10.1111/j.1095-8339.2009.01002.x.

Christie PJ, Alfenito MR & Walbot V (1994) Impact of low-temperature stress on general phenylpropanoid and anthocyanin pathways: Enhancement of transcript abundance and anthocyanin pigmentation in maize seedlings. *Planta*, **194**(4): 541-549. doi:10.1007/BF00714468.

Coen ES & Meyerowitz EM (1991) The war of the whorls: genetic interactions controlling flower development. *Nature*, **353**(6339): 31-37. doi:10.1038/353031a0.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA & Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science*, **298**(5597): 1415-1418. doi:10.1126/science.1077769.

Dalziel AC, Rogers SM & Schulte PM (2009) Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology*, **18**(24): 4997-5017. doi:10.1111/j.1365-294X.2009.04427.x.

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

Darwin CR (1868) *The Variation of Animals and Plants Under Domestication*. John Murray, London, UK

Devey DS, Bateman RM, Fay MF & Hawkins JA (2008) Friends or relatives? Phylogenetics and species delimitation in the controversial European orchid genus *Ophrys*. *Annals of Botany*, **101**(3): 385-402. doi:10.1093/aob/mcm299.

Dixon RA & Paiva NL (1995) Stress-induced phenylpropanoid metabolism. *The Plant Cell*, **7**(7): 1085-1097. doi:10.1105/tpc.7.7.1085.

Edwards MD, Stuber CW & Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics*, **116**(1): 113-125.

Falcone Ferreyra ML, Rius SP & Casati P (2012) Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Frontiers in Plant Science*, **3**: 222. doi:10.3389/fpls.2012.00222.

Forkmann G & Dangelmayr B (1980) Genetic control of chalcone isomerase activity in flowers of Dianthus caryophyllus. *Biochemical Genetics*, **18**(5): 519-527. doi:10.1007/BF00484399.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R*, et al* (2010) Meta-analysis increases to 71 the tally of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, **42**(12): 1118-1125. doi:10.1038/ng.717.

Friedman J, Twyford AD, Willis JH & Blackman BK (2015) The extent and genetic basis of phenotypic divergence in life history traits in *Mimulus guttatus*. *Molecular Ecology*, **24**(1): 111-122. doi:10.1111/mec.13004.

Gausman HW (1983) Visible light reflectance, transmittance, and absorbance of differently pigmented cotton leaves. *Remote Sensing of Environment*, **13**. doi:10.1016/0034-4257(83)90041-x.

Gilbert RI (1971) An unusual anthocyanin in *Antirrhinum majus*. *Phytochemistry*, **10**(11): 2848-2849. doi:10.1016/S0031-9422(00)97309-6.

Glover BJ (2014) *Understanding Flowers and Flowering: an Integrated Approach*. Oxford University Press, Oxford, UK, ISBN: 9780199661596.

Glover BJ & Martin CR (2012) Anthocyanins. *Current Biology*, **22**(5): R147-R150. doi:10.1016/j.cub.2012.01.021.

Hansen DM, Van der Niet T & Johnson SD (2012) Floral signposts: testing the significance of visual 'nectar guides' for pollinator behaviour and plant fitness. *Proceedings of the Royal Society B: Biological Sciences*, **279**(1729): 634-639. doi:10.1098/rspb.2011.1349.

Hashimoto H, Uragami C & Cogdell RJ (2016) Carotenoids and photosynthesis. *Carotenoids in Nature: Biosynthesis, Regulation and Function*, Stange C (Ed), Springer International Publishing, Cham: 111-139.

Hewitt GM (1988) Hybrid zones: natural laboratories for evolutionary studies. *Trends in Ecology & Evolution*, **3**(7): 158-167. doi:10.1016/0169-5347(88)90033-X.

Hubby JL & Lewontin RC (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, **54**(2): 577-594.

Hudson A, Critchley J & Erasmus Y (2008) The genus *Antirrhinum* (snapdragon): a flowering plant model for evolution and development. *Cold Spring Harbor Protocols*, **2008**(10): pdb.emo100. doi:10.1101/pdb.emo100.

Jiao K, Li X, Guo W, Su S & Luo D (2017) High-throughput RNA-seq data analysis of the single nucleotide polymorphisms (SNPs) and zygomorphic flower development in pea (*Pisum sativum* L.). *International Journal of Molecular Sciences*, **18**(12): 2710. doi:10.3390/ijms18122710.

Jiggins CD, Wallbank RWR & Hanly JJ (2017) Waiting in the wings: what can we learn about gene co-option from the diversification of butterfly wing patterns? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**(1713). doi:10.1098/rstb.2015.0485.

Johnston SE, Gratten J, Berenos C, Pilkington JG, Clutton-Brock TH, Pemberton JM & Slate J (2013) Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature*, **502**: 93. doi:10.1038/nature12489

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L*, et al* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**: 203. doi:10.1038/nature10341

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P & Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**(12): 1647-1649. doi:10.1093/bioinformatics/bts199.

Kerem B, Rommens J, Buchanan J, Markiewicz D, Cox T, Chakravarti A, Buchwald M & Tsui L (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**(4922): 1073-1080. doi:10.1126/science.2570460.

King EG & Long AD (2017) The Beavis Effect in next-generation mapping panels in *Drosophila melanogaster*. *G3: Genes|Genomes|Genetics*, **7**(6): 1643-1652. doi:10.1534/g3.117.041426.

Kirkpatrick M (2010) How and why chromosome inversions evolve. *PLOS Biology*, **8**(9): e1000501. doi:10.1371/journal.pbio.1000501.

Kirkpatrick M & Barton NH (2006) Chromosome inversions, local adaptation and

speciation. *Genetics*, **173**(1): 419-434. doi:10.1534/genetics.105.047985.

Koes RE, Quattrocchio F & Mol JNM (1994) The flavonoid biosynthetic pathway in plants: function and evolution. *BioEssays*, **16**(2): 123-132. doi:doi:10.1002/bies.950160209.

Koressaar T & Remm M (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**(10): 1289-1291. doi:10.1093/bioinformatics/btm091.

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, **304**: 412. doi:10.1038/304412a0.

Krimbas CB & Powell JR (1992) Introduction. *Drosophila Inversion Polymorphism*, Krimbas CB and Powell JR (Eds), CRC Press, Boca Raton, FL, US.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C*, et al* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, **40**(5): 584-591. doi:10.1038/ng.125.

Levy SE & Myers RM (2016) Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, **17**(1): 95-115. doi:10.1146/annurev-genom-083115-022413.

LGC Ltd (2013) KASP genotyping chemistry. Retrieved 23 January 2018, 2018, from http://www.lgcgroup.com/LGCGroup/media/PDFs/Products/Genotyping/KASP-genotyping-chemistry-User-guide.pdf?ext=.pdf.

Li H (2013) bwa - Burrows-Wheeler Alignment Tool. Wellcome Sanger Institute, Hinxton, UK. Available at http://bio-bwa.sourceforge.net.

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

Liu L, White MJ & MacRae TH (1999) Transcription factors and their genes in higher plants. *European Journal of Biochemistry*, **262**(2): 247-257. doi:10.1046/j.1432-1327.1999.00349.x.

Lopes-da-Silva F, Escribano-Baion MT & Santos-Buelga C (2007) Stability of pelargonidin 3-glucoside in model solutions in the presence and absence of flavanols.

American Journal of Food Technology, **2**(7): 602-617. doi:10.3923/ajft.2007.602.617.

Magwene PM, Willis JH & Kelly JK (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLOS Computational Biology*, **7**(11): e1002255. doi:10.1371/journal.pcbi.1002255.

Mansfeld BN & Grumet R (2017) QTLseqr: An R package for bulk segregant analysis with next-generation sequencing. *bioRxiv*. doi:10.1101/208140.

Martin C & Paz-Ares J (1997) MYB transcription factors in plants. *Trends in Genetics*, **13**(2): 67-73. doi:10.1016/S0168-9525(96)10049-4.

Martin CR, 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**(7): 1625-1630. doi:10.1002/j.1460-2075.1985.tb03829.x.

Martin CR, Prescott A, Mackay S, Bartlett J & Vrijlandt E (1991) Control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. *Plant Journal*, **1**(1): 37-49. doi:10.1111/j.1365-313X.1991.00037.x.

Mather K (1947) Species crosses in *Antirrhinum* 1: genetic isolation of the species *majus*, *glutinosum* and *orontium*. *Heredity*, **1**(2): 175-&. doi:10.1038/hdy.1947.12.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M & DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9): 1297-1303. doi:10.1101/gr.107524.110.

Meshi T & Iwabuchi M (1995) Plant transcription factors. *Plant and Cell Physiology*, **36**(8): 1405-1420.

Michelmore RW, Paran I & Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(21): 9828-9832. doi:10.1073/pnas.88.21.9828.

Miki S, Wada KC & Takeno K (2015) A possible role of an anthocyanin filter in low-intensity light stress-induced flowering in Perilla frutescens var. crispa. *Journal of Plant Physiology*, **175**: 157-162. doi:10.1016/j.jplph.2014.12.002.

Moyroud E, Wenzel T, Middleton R, Rudall PJ, Banks H, Reed A, Mellers G, Killoran P, Westwood MM, Steiner U, Vignolini S & Glover BJ (2017) Disorder in convergent floral nanostructures enhances signalling to bees. *Nature*, **550**: 469. doi:10.1038/nature24285.

Nakayama T (2002) Enzymology of aurone biosynthesis. *Journal of Bioscience and Bioengineering*, **94**(6): 487-491. doi:10.1016/S1389-1723(02)80184-0.

Nakayama T, Yonekura-Sakakibara K, Sato T, Kikuchi S, Fukui Y, Fukuchi-Mizutani M, Ueda T, Nakao M, Tanaka Y, Kusumi T & Nishino T (2000) Aureusidin synthase: a polyphenol oxidase homolog responsible for flower coloration. *Science*, **290**(5494): 1163-1166. doi:10.1126/science.290.5494.1163.

NCBI Resource Coordinators (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **45**(D1): D12-D17. doi:10.1093/nar/gkw1071.

Nisar N, Li L, Lu S, Khin Nay C & Pogson Barry J (2015) Carotenoid metabolism in plants. *Molecular Plant*, **8**(1): 68-82. doi:https://doi.org/10.1016/j.molp.2014.12.007.

Noda K, Glover BJ, Linstead P & Martin C (1994) Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor. *Nature*, **369**: 661. doi:10.1038/369661a0.

Ono E, Fukuchi-Mizutani M, Nakamura N, Fukui Y, Yonekura-Sakakibara K, Yamaguchi M, Nakayama T, Tanaka T, Kusumi T & Tanaka Y (2006) Yellow flowers generated by expression of the aurone biosynthetic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(29): 11075-11080. doi:10.1073/pnas.0604246103.

Osorio D & Vorobyev M (2008) A review of the evolution of animal colour vision and visual communication signals. *Vision Research*, **48**(20): 2042-2051. doi:https://doi.org/10.1016/j.visres.2008.06.018.

Owen CR & Bradshaw HD (2011) Induced mutations affecting pollinator choice in *Mimulus lewisii* (Phrymaceae). *Arthropod-Plant Interactions*, **5**(3): 235. doi:10.1007/s11829-011-9133-8.

Park ST & Kim J (2016) Trends in next-generation sequencing and a new era for whole genome sequencing. *International Neurourology Journal*, **20**(Suppl 2): S76-83. doi:10.5213/inj.1632742.371.

Parts L, Cubillos FA, Warringer J, Jain K, Salinas F, Bumpstead SJ, Molin M, Zia A, Simpson JT, Quail MA, Moses A, Louis EJ, Durbin R & Liti G (2011) Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, **21**(7): 1131-1138. doi:10.1101/gr.116731.110.

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S & Pérez-Enciso M (2012) SNP calling by sequencing pooled samples. *BMC Bioinformatics*, **13**(1): 1-8. doi:10.1186/1471-2105-13-239.

Rausher MD (2006) The evolution of flavonoids and their genes. *The Science of Flavonoids*, Grotewold E (Ed), Springer Science + Business Media, Inc, New York, NY, US.

Reed RD, Papa R, Martin A, Hines HM, Counterman BA, Pardo-Diaz C, Jiggins CD, Chamberlain NL, Kronforst MR, Chen R, Halder G, Nijhout HF & McMillan WO (2011) *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, **333**(6046): 1137-1141. doi:10.1126/science.1208227.

Rieseberg LH, Whitton J & Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**(2): 713-727.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G & Mesirov JP (2011) Integrative Genomics Viewer. *Nature Biotechnology*, **29**: 24. doi:10.1038/nbt.1754

Rothmaler W (1956) *Taxonomische monographie der gattung Antirrhinum*. Akademie-Verlag, Berlin

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H & Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**(4732): 1350-1354. doi:10.1126/science.2999980.

Saito N & Harborne JB (1992) Correlations between anthocyanin type, pollinator and flower colour in the Labiatae. *Phytochemistry*, **31**(9): 3009-3015. doi:10.1016/0031-9422(92)83437-4.

Schiestl FP & Johnson SD (2013) Pollinator-mediated evolution of floral signals. *Trends in Ecology & Evolution*, **28**(5): 307-315. doi:10.1016/j.tree.2013.01.019.

Schwarz-Sommer Z, Davies B & Hudson A (2003) An everlasting pioneer: the story of Antirrhinum research. *Nature Reviews Genetics*, **4**: 655. doi:10.1038/nrg1127.

Schwinn KE, Venail J, Shang Y, Mackay S, Alm V, Butelli E, Oyama RK, Bailey PC, Davies K & Martin CR (2006) A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. *Plant Cell*, **18**(4): 831-851. doi:10.1105/tpc.105.039255.

Scott-Moncrieff R (1930) Natural anthocyanin pigments: the magenta flower pigment of Antirrhinum majus. *Biochemical Journal*, **24**(3): 753-766.

Shang Y, Venail J, Mackay S, Bailey PC, Schwinn KE, Jameson PE, Martin CR & Davies KM (2011) The molecular basis for venation patterning of pigmentation and its effect on pollinator attraction in flowers of *Antirrhinum*. *New Phytologist*, **189**(2): 602-615. doi:10.1111/j.1469-8137.2010.03498.x.

Simeone P & Alberti S (2014) Epigenetic heredity of human height. *Physiological Reports*, **2**(6): e12047. doi:10.14814/phy2.12047.

Skaar I, Adaku C, Jordheim M, Byamukama R, Kiremire B & Andersen ØM (2014) Purple anthocyanin colouration on lower (abaxial) leaf surface of Hemigraphis colorata (Acanthaceae). *Phytochemistry*, **105**: 141-146. doi:10.1016/j.phytochem.2014.05.016.

Sobel J & Streisfeld M (2013) Flower color as a model system for studies of plant evo-devo. *Frontiers in Plant Science*, **4**(321). doi:10.3389/fpls.2013.00321.

Sommer H & Saedler H (1986) Structure of the chalcone synthase gene of *Antirrhinum majus*. *Molecular and General Genetics MGG*, **202**(3): 429-434. doi:10.1007/BF00333273.

Song J, Li Z, Liu Z, Guo Y & Qiu L-J (2017) Next-generation sequencing from bulked-segregant analysis accelerates the simultaneous identification of two qualitative genes in soybean. *Frontiers in Plant Science*, **8**: 919. doi:10.3389/fpls.2017.00919.

Stankowski S, Sobel JM & Streisfeld MA (2015) The geography of divergence with gene flow facilitates multitrait adaptation and the evolution of pollinator isolation in Mimulus aurantiacus. *Evolution*, **69**(12): 3054-3068. doi:10.1111/evo.12807.

Stanton KA, Edger PP, Puzey JR, Kinser T, Cheng P, Vernon DM, Forsthoefel NR

& Cooley AM (2017) A whole-transcriptome approach to evaluating reference genes for quantitative gene expression studies: a case study in *Mimulus*. *G3: Genes|Genomes|Genetics*, **7**(4): 1085.

Stern DL & Orgogozo V (2009) Is genetic evolution predictable? *Science*, **323**(5915): 746-751. doi:10.1126/science.1158997.

Stickland RG & Harrison BJ (1977) Precursors and genetic control of pigmentation. *Heredity*, **39**: 327. doi:10.1038/hdy.1977.73.

Stommel JR, Lightbourn GJ, Winkel BS & Griesbach RJ (2009) Transcription factor families regulate the anthocyanin biosynthetic pathway in *Capsicum annuum*. *Journal of the American Society for Horticultural Science*, **134**(2): 244-251.

Strack D, Vogt T & Schliemann W (2003) Recent advances in betalain research. *Phytochemistry*, **62**(3): 247-269. doi:https://doi.org/10.1016/S0031-9422(02)00564-2.

Streisfeld MA & Rausher MD (2011) Population genetics, pleiotropy, and the preferential fixation of mutations during adaptive evolution. *Evolution*, **65**(3): 629-642. doi:10.1111/j.1558-5646.2010.01165.x.

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, **14**(1): 43-59. doi:10.1002/jez.1400140104.

Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano LM, Kamoun S & Terauchi R (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*, **74**(1): 174-183. doi:10.1111/tpj.12105.

Tanaka Y, Sasaki N & Ohmiya A (2008) Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *The Plant Journal*, **54**(4): 733-749. doi:10.1111/j.1365-313X.2008.03447.x.

Tastard E, Ferdy JB, Burrus M, Thébaud C & Andalo C (2012) Patterns of floral colour neighbourhood and their effects on female reproductive success in an *Antirrhinum* hybrid zone. *Journal of Evolutionary Biology*, **25**(2): 388-399. doi:10.1111/j.1420-9101.2011.02433.x.

Tavares H (2014) Evolutionary genetics and genomics of flower colour loci in an

Antirrhinum hybrid zone. University of East Anglia, **PhD**.

Tavares H, Whibley A, Field D, Bradley D, Couchman M, Copsey L, Elleouet J, Burrus M, Andalo C, Li M*, et al* (in review) Selection and gene flow shape genomic islands that control floral guides.

Turner LM & Harr B (2014) Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *eLife*, **3**: e02504. doi:10.7554/eLife.02504.

Twyford AD & Friedman J (2015) Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution*, **69**(6): 1476-1486. doi:doi:10.1111/evo.12663.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth B, Remm M & Rozen S (2012) Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, **40**(15): e115.

Valadon LRG & Mummery RS (1968) Carotenoids in floral parts of a narcissus, a daffodil and a tulip. *Biochemical Journal*, **106**(2): 479-484.

Vargas P, Carrió E, Guzmán B, Amat E & Güemes J (2009) A geographical pattern of *Antirrhinum* (Scrophulariaceae) speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. *Journal of Biogeography*, **36**(7): 1297-1312. doi:10.1111/j.1365-2699.2008.02059.x.

Vargas P, Ornosa C, Ortiz-Sánchez FJ & Arroyo J (2010) Is the occluded corolla of *Antirrhinum* bee-specialized? *Journal of Natural History*, **44**(23-24): 1427-1443. doi:10.1080/00222930903383552.

Venail J (2005) Floral pigmentation in the genus *Antirrhinum*. University of East Anglia, Norwich, **PhD**.

Vignolini S, Moyroud E, Hingant T, Banks H, Rudall Paula J, Steiner U & Glover BJ (2014) The flower of *Hibiscus trionum* is both visibly and measurably iridescent. *New Phytologist*, **205**(1): 97-101. doi:10.1111/nph.12958.

Wallbank RWR, Baxter SW, Pardo-Diaz C, Hanly JJ, Martin SH, Mallet J, Dasmahapatra KK, Salazar C, Joron M, Nadeau N, McMillan WO & Jiggins CD (2016) Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLOS Biology*, **14**(1): e1002353. doi:10.1371/journal.pbio.1002353.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy

RM, Perry JRB, Stevens S, Hall AS*, et al* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, **40**(5): 575-583. doi:10.1038/ng.121.

Whibley AC (2004) Molecular and genetic variation underlying the evolution of flower colour in *Antirrhinum*. University of East Anglia, Norwich, **PhD**.

Whibley AC, Langlade NB, Andalo C, Hanna AI, Bangham A, Thébaud C & Coen ES (2006) Evolutionary paths underlying flower color variation in *Antirrhinum*. *Science*, **313**(5789): 963-966. doi:10.1126/science.1129161.

Wickam H (2009) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY, US. Available at http://ggplot2.org.

Willis KJ (Ed) (2017). *State of the World's Plants 2017*. Royal Botanic Gardens, Kew, London, UK.

Wilson P, Castellanos MC, Hogue JN, Thomson JD & Armbruster WS (2004) A multivariate search for pollination syndromes among penstemons. *Oikos*, **104**(2): 345-361. doi:10.1111/j.0030-1299.2004.12819.x.

Wilson Y & Hudson A (2011) The evolutionary history of *Antirrhinum* suggests that ancestral phenotype combinations survived repeated hybridizations. *The Plant Journal*, **66**(6): 1032-1043. doi:10.1111/j.1365-313X.2011.04563.x.

Wu CA, Streisfeld MA, Nutter LI & Cross KA (2013) The genetic basis of a rare flower color polymorphism in *Mimulus lewisii* provides insight into the repeatability of evolution. *PLOS One*, **8**(12): e81173. doi:10.1371/journal.pone.0081173.

Xue Y *et al* (unpublished *Antirrhinum* genome; Beijing Institute of Genomics and the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences).

Yuan Y-W, Rebocho AB, Sagawa JM, Stanley L & Bradshaw HD (2016) Competition between anthocyanin and flavonol biosynthesis produces spatial pattern variation of floral pigments between Mimulus species. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(9): 2448-2453. doi:10.1073/pnas.1515294113.

Yuan Y-W, Sagawa JM, Di Stilio VS & Bradshaw HD (2013) Bulk segregant analysis of an induced floral mutant identifies a *mixta*-like r2r3 *MYB* controlling nectar guide formation in *Mimulus lewisii*. *Genetics*, **194**(2): 523-528. doi:10.1534/genetics.113.151225.

Yuan YW, Sagawa JM, Frost L, Vela JP & Bradshaw HD, Jr. (2014) Transcriptional control of floral anthocyanin pigmentation in monkeyflowers (Mimulus). *New Phytologist*, **204**(4): 1013-1027. doi:10.1111/nph.12968.

Zhang B, Liu C, Wang Y, Yao X, Wang F, Wu J, King GJ & Liu K (2015) Disruption of a *CAROTENOID CLEAVAGE DIOXYGENASE 4* gene converts flower colour from white to yellow in Brassica species. *New Phytologist*, **206**(4): 1513-1526. doi:10.1111/nph.13335.

Zhao D & Tao J (2015) Recent advances on the development and regulation of flower color in ornamental plants. *Frontiers in Plant Science*, **6**(261). doi:10.3389/fpls.2015.00261.

INDEX

8 Appendices

8.1 **Appendix 1** Primers used for genotyping *ROSEA* and *ELUTA*

Set number	Primer number	Description	Orientation	Focal SNP	Sequence
1	1	*ROS1* promoter (*pseudomajus* VIC)	F	541737	GAAGGTCGGAGTCAACGGATTGGGCATAGTACGTATTAAACGC
	2	*ROS1* promoter (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTGGGCATAGTACGTATTAAACGA
	3	*ROS1* promoter (common reverse)	R		GGTCCAAGTACCTTTTCTCACT
2	4	*ROS2* exon 3 (*pseudomajus* VIC)	F	566868	GAAGGTCGGAGTCAACGGATTAAATAGTAAAGAAACTAATATC
	5	*ROS2* exon 3 majus FAM	F		GAAGGTGACCAAGTTCATGCTAAATAGTAAAGAAACTAATATT
	6	*ROS2* exon 3 (common reverse)	R		CGTGCAATCCATTGAAGGTCCG
3	7	*ROS3* exon 1 (*pseudomajus* VIC)	F	573874	GAAGGTCGGAGTCAACGGATTTGACGCAATGCGTGGAGAAGTT
	8	*ROS3* exon 1 (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTTGACGCAATGCGTGGAGAAGTA
	9	*ROS3* exon 1 (common reverse)	R		CCTGCTCTGAGCGGGACTTGAT
4	10	*ROS3* exon 3 (*pseudomajus* VIC)	F	576036	GAAGGTCGGAGTCAACGGATTCTTGTCCAAATTGCATGAAACA
	11	*ROS3* exon 3 (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTCTTGTCCAAATTGCATGAAACT
	12	*ROS3* exon 3 (common reverse)	R		GTTCGCTTCTCTCACTTCATTT
5	13	*ROS3* exon 3 (*pseudomajus* VIC)	F	576237	GAAGGTCGGAGTCAACGGATTTCATCGTGTTTTCTCCATCGAT
	14	*ROS3* exon 3 (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTTCATCGTGTTTTCTCCATCGAC
	15	*ROS3* exon 3 (common reverse)	R		GATCGTCCATGTCTACCACGTC
6	16	*EL* exon 3 (*pseudomajus* VIC)	F	714767	GAAGGTCGGAGTCAACGGATTGAAGCCGTTAAGTCGCAGGTGC
	17	*EL* exon 3 (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTGAAGCCGTTAAGTCGCAGGTGT
	18	*EL* exon 3 (common reverse)	R		GTTCCCCTGTTGATCCTGAAGA
7	19	Downstream *ROS3* (*pseudomajus* VIC)	F	578688	GAAGGTCGGAGTCAACGGATTAAATGCGTACAATTCTAATATG
	20	Downstream *ROS3* (JI7 FAM)	F		GAAGGTGACCAAGTTCATGCTAAATGCGTACAATTCTAATATC
	21	Downstream *ROS3* (common reverse)	R		CCAATATAACAACTTGATGGCC
8	22	*ROS2* exon 3 (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTGCACAATTTGTTGTTTTCTAAC

	23	*ROS2* exon 3 (JI7 FAM)	F	566671	GAAGGTGACCAAGTTCATGCTGCACAATTTGTTGTTTTCTAAT
	24	*ROS2* exon 3 (common reverse)	R		CCACGCCTAAATTCTTCCCCA
9	25	*ROS2* exon 3 (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTGTTTTTACCGTTAATGATTGAC
	26	*ROS2* exon 3 (JI7 FAM)	F	566713	GAAGGTGACCAAGTTCATGCTGTTTTTACCGTTAATGATTGAT
	27	*ROS2* exon 3 (common reverse)	R		CGTTCTCCATCCACGCCTAA
10	28	*ROS1* intron (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTCAGTTGACACTTTATCTTGGAC
	29	*ROS1* intron (JI7 FAM)	F	543354	GAAGGTGACCAAGTTCATGCTCAGTTGACACTTTATCTTGGAT
	30	*ROS1* intron (common reverse)	R		GAGTTTCAACAAGACGGGAGC
11	31	*ROS1* promoter (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTCCTGGCTCCACCCTATGATGG
	32	*ROS1* promoter (JI7 FAM)	F	541203	GAAGGTGACCAAGTTCATGCTTCCTGGCTCCACCCTATGATGT
	33	*ROS1* promoter (common reverse)	R		TCCTTAATCATCTGTCCTTTCATTTCA
12	34	Downstream *ROS3* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTATAATTTTGATAATGATATAA
	35	Downstream *ROS3* (JI7 FAM)	F	578296	GAAGGTGACCAAGTTCATGCTTATAATTTTGATAATGATATAG
	36	Downstream *ROS3* (common reverse)	R		AAAGTTGGCAACCAGTTAGCT
13	37	Downstream *ROS3* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTATTCACATTAGTTGTTATTTT
	38	Downstream *ROS3* (JI7 FAM)	F	578680	GAAGGTGACCAAGTTCATGCTTATTCACATTAGTTGTTATTTG
	39	Downstream *ROS3* (common reverse)	R		ACGTCTAACTTGACTTCAAAAATAGT
14	40	Downstream *ROS3* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTCAGCTATAGTTATGGATTTTCG
	41	Downstream *ROS3* (JI7 FAM)	F	597501	GAAGGTGACCAAGTTCATGCTCAGCTATAGTTATGGATTTTCC
	42	Downstream *ROS3* (common reverse)	R		CTGCAAAAGAGTTGACTGAGGC
15	43	Downstream of *EL* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTAATTAATTCCTATAATTTCAAA
	44	Downstream of *EL* (JI7 FAM)	F	711121	GAAGGTGACCAAGTTCATGCTAATTAATTCCTATAATTTCAAC
	45	Downstream of *EL* (common reverse)	R		AGCGAAGGTCTAGTCCACTT
16	46	Downstream of *EL* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTCTCTTTTCCTTTTGATAAGATC
	47	Downstream of *EL* (JI7 FAM)	F	714357	GAAGGTGACCAAGTTCATGCTCTCTTTTCCTTTTGATAAGATT

	48	Downstream of *EL* (common reverse)	R		TCCTTGTGGTCTCTCTTTTCGT
	49	*EL* exon 3 (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTGAAGCCGTTAAGTCGCAGGTGC
17	50	*EL* exon 3 (JI7 FAM)	F	714767	GAAGGTGACCAAGTTCATGCTGAAGCCGTTAAGTCGCAGGTGT
	51	*EL* exon 3 (common reverse)	R		TTCGGACAATCATTCCTCGGA
	52	*EL* exon 3 (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTGATCGACGCCCATGTTTATCAA
18	53	*EL* exon 3 (JI7 FAM)	F	714954	GAAGGTGACCAAGTTCATGCTGATCGACGCCCATGTTTATCAG
	54	*EL* exon 3 (common reverse)	R		CGGGGCGAATGGATGATGAA
	55	*EL* intron 1 (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTAGAAGAAATCAGGCTTGCATGT
19	56	*EL* intron 1 (JI7 FAM)	F	716191	GAAGGTGACCAAGTTCATGCTAGAAGAAATCAGGCTTGCATGA
	57	*EL* intron 1 (common reverse)	R		ATGGTGAAGGATGTTGGCGT
	58	Upstream of *ROS1* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTCGGTCATCTACCAAGGAACTGG
20	59	Upstream of *ROS1* (JI7 FAM)	F	156123	GAAGGTGACCAAGTTCATGCTCGGTCATCTACCAAGGAACTGA
	60	Upstream of *ROS1* (common reverse)	R		GTTGCCACTAAACCACTGGCCTG
	61	Upstream of *ROS1* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTGGTGGTTTTATTTGGTTATAA
21	62	Upstream of *ROS1* (JI7 FAM)	F	312030	GAAGGTGACCAAGTTCATGCTTGGTGGTTTTATTTGGTTATAC
	63	Upstream of *ROS1* (common reverse)	R		GAATATCATGCCATTTGCATCC
	64	Upstream of *ROS1* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTTCAATTATTTGATAAGAGATA
22	65	Upstream of *ROS1* (JI7 FAM)	F	313548	GAAGGTGACCAAGTTCATGCTTTCAATTATTTGATAAGAGATG
	66	Upstream of *ROS1* (common reverse)	R		GTTTGGCGTAACAATTGTTTGG
	67	Upstream of *ROS1* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTTGAATAAAAACCTGCTGGCCTA
23	68	Upstream of *ROS1* (JI7 FAM)	F	394415	GAAGGTGACCAAGTTCATGCTTGAATAAAAACCTGCTGGCCTG
	69	Upstream of *ROS1* (common reverse)	R		GAACATATCCTGCATTAATCAA
	70	Upstream of *ROS1* (*pseudomajus* VIC)	F		GAAGGTCGGAGTCAACGGATTATCGTTGACACAGTAAAACTGT
24	71	Upstream of *ROS1* (JI7 FAM)	F	476066	GAAGGTGACCAAGTTCATGCTATCGTTGACACAGTAAAACTGA
	72	Upstream of *ROS1* (common reverse)	R		GTTTGTTATTAACAGAACCATT

	73	*ELUTA* (*pseudomajus* VIC)	R		GAAGGTCGGAGTCAACGGATTGGCGTCGATCCAAATAACCACC
25	74	*ELUTA* (JI7 FAM)	R	714921	GAAGGTGACCAAGTTCATGCTGGCGTCGATCCAAATAACCACT
	75	*ELUTA* (common forward)	F		AGACAGACAATGTTGTAATATC