Genetic basis of flower colour as a model for adaptive evolution

by

Lenka Matejovičová

January, 2022

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Eva Benková, Chair
Nick Barton
Beatriz Vicoso
Magnus Nordborg

Institute of Science and Technology

The thesis of Lenka Matejovičová, titled *Genetic basis of flower colour as a model for adaptive evolution*, is approved by:

Supervisor: Nick Barton, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Beatriz Vicoso, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Magnus Nordborg, Gregor Mendel Institute of molecular Plant Biology, Vienna, Austria

Signature: _____

Defense Chair: Eva Benková, IST Austria, Klosterneuburg, Austria

Signature: _____

Signed page is on file

IST Austria Thesis, ISSN: 2663-337X

ISBN: 978-3-99078-016-9

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Lenka Matejovičová
January, 2022

Signed page is on file

Abstract

Although we often see studies focusing on simple or even discrete traits in studies of colouration, the variation of "appearance" phenotypes found in nature is often more complex, continuous and high-dimensional. Therefore, we developed automated methods suitable for large datasets of genomes and images, striving to account for their complex nature, while minimising human bias. We used these methods on a dataset of more than $20,000$ plant SNP genomes and corresponding flower images from a hybrid zone of two subspecies of *Antirrhinum majus* with distinctly coloured flowers to improve our understanding of the genetic nature of the flower colour in our study system.

Firstly, we use the advantage of large numbers of genotyped plants to estimate the haplotypes in the main flower colour regulating region. We study colour- and geography-related characteristics of the estimated haplotypes and how they connect to their relatedness. We show discrepancies from the expected flower colour distributions given the genotype and identify particular haplotypes leading to unexpected phenotypes. We also confirm a significant deficit of the double recessive recombinant and quite surprisingly, we show that haplotypes of the most frequent parental type are much less variable than others.

Secondly, we introduce our pipeline capable of processing tens of thousands of full flower images without human interaction and summarising each image into a set of informative scores. We show the compatibility of these machine-measured flower colour scores with the previously used manual scores and study impact of external effect on the resulting scores. Finally, we use the machine-measured flower colour scores to fit and examine a phenotype cline across the hybrid zone in Planoles using full flower images as opposed to discrete, manual scores and compare it with the genotypic cline.

Acknowledgements

Here, I would like to thank people who helped me to write this thesis. First and foremost, I would like to thank my supervisor **Nick Barton** for taking me on this project, for his guidance, patience and support and to my internal committee member **Beatriz Vicoso** for her words of wisdom and encouragement.

I would like to thank to members of the Barton group, especially to **David Field**, **Maria Melo** and **Melinda Pickup** for introducing me to the colourful world of *Antirrhinum* biology and the field work and to **Louise Arathoon**, **Åshild Dybdal**, **Arka Pal**, **Daria Shipilina**, **Sean Stankowski** and **Parvathy Surendranadh** for sharing their knowledge and helping me with data acquisition. I would also like to express my gratitude to the rest of the *Antirrhinum* team, including hundreds of volunteers, all working together to build together the main *Antirrhinum* dataset and, of course, to the group assistant **Astrid Bonventre-Darthe**, who made sure that the field work ran smoothly.

I found IST Austria to be a great place to do research, especially due to its excellent Scientific Service Units. In particular, I am very thankful to **Alois Schlögl** and **Stephan Stadlbauer** from Scientific Computing, and **Christoph Sommer** from Imaging and Optics facility, who introduced me to image analysis.

Throughout my stay at the Institute, I was very lucky to meet lovely people and I would like to thank **Julia Asimakis**, **Marta Dravecká**, **Enikő Edelsbrunner**, **Zuzana Masárová**, **Simon Mayer**, **Thomas Moser**, **Georg Osang**, **Hana Semerádová**, **Tomáš Skřivan**, **Josef Tkadlec**, **Viktor Toman**, **Anja Westram** and many others for being there for me when the times got tough.

Last but not least, I would like to thank my beloved family, who raised me up and who have cared for me, inspired me and believed in me ever since.

About the Author

Lenka Matejovičová completed a BSc in Applied Mathematics at Comenius University in Bratislava, Slovakia and an MPhil in Computational Biology at the University of Cambridge before coming to IST Austria, and joining the Barton group in January 2017. Her goal in research is to use Data Analysis, Computational Biology, Bioinformatics and large datasets to solve problems stemming from Life Sciences and to support the dialogue between Mathematicians and Biologists. During her PhD studies she also organised and co-taught an introductory Mathematics course for Life Scientists and received a "Golden Sponge Award" for the best teaching assistant of the year.

Table of Contents

CHAPTER 1

Introduction

Lately, *Big Data* is turning from a fad into an imperative in the Life Sciences. In pursuit of holistic views, fashionably small p-values and novel effects of ever decreasing magnitude, biologists have struggled to make good use of ambitious and expensive datasets.

Evolutionary biology is no exception to this trend. Although evolution shapes the variation and diversity of everything that is alive, it acts in such a gradual and long-term way, that one lifetime is often too short and one's senses too limited to witness any significant effects (let alone in the duration of one PhD). Therefore, considerable statistical power is necessary to illuminate its workings and in particular, large-scale datasets and corresponding analytical methods are needed. Hopefully, this thesis will serve as a good example.

To support any evolutionary theory with data, one needs to have a good understanding of the phenotype and its relationship with underlying genotype, as well as their connection to fitness. In this thesis, we argue with an example, that using objective and scalable methods on large datasets helps us to better understand both the genotype and the phenotype. But most importantly, it helps us to broaden our understanding of the relationship between them, while setting the scene for investigating their connection to fitness.

1.1 Quantitative approaches to phenotype

We have been used to think about Mendelian traits (i.e. simple traits governed by one or two loci) as discrete and easily assigned into a handful of phenotypic categories. In particular, these traits were contrasted with continuous quantitative traits such as height, affected by large numbers of loci. However, Mendelian traits do in fact exhibit continuous variation, which is particularly noticeable in visible phenotypes, such as colouration.

For example, in [13] the authors noted that although the shell colour of the land snail *Cepea nemoralis* is commonly assigned one of three categories (yellow, pink or brown), in practice, for many shells it is difficult to assign a single colour category. Using continuous reflectance measures they found that although the shell colours do cluster into three groups corresponding to the human-defined colour categories, shell colour variation in fact *is continuous*.

The quest to understand continuous, high-dimensional phenotypes in terms of Mendelian traits has been addressed in several systems, typically using automated continuous phenotype quantification methods. While the above example used reflectance measured at several areas

of the shell, it is also quite common to use digital imaging (photographs), especially when the colour pattern is also of interest.

For example, in [77] the authors developed an eye colour quantification pipeline that identifies the distribution of the two pigments and non-pigmented areas throughout the human iris. This allowed them to quantify human eye colour much more accurately and objectively than the previous manual scoring systems. Therefore they suggested the use of the new data in genome-wide association studies, possibly identifying loci which they would overlook using the limited, human-scored dataset.

Similarly, although the presence of two types of pigments on the throat of an Australian lizard *Ctenophorus decresii* is mostly Mendelian and can be categorised into four non-overlapping groups, the intensity and distribution of the pigments still varies greatly within each group [49]. Therefore, the authors used automated phenotype quantification from the photographs to better understand the genetic basis of the polymorphism, evaluated different Mendelian models of inheritance of several quantitatively characterised traits, and calculated their heritability using a pedigree.

Many appearance phenotypes are in fact continuous and multidimensional, although mostly Mendelian in nature. To understand them in their complexity, we would be hindered by limited, human-designed and possibly biased discrete phenotype-scoring schemes. However, we are not dependent on manual scores, because we can take advantage of objective quantitative techniques, photographs, or even 3D models that capture phenotypes in their entirety.

1.2 Hybrid zones

When looking for a good source of data to investigate questions of evolution, one can, of course, turn to controlled experiments in a laboratory, or in a greenhouse. Experimental evolution has advantages, especially as it is under the control of the researcher throughout the duration of the experiment. On the other hand, due to limited resources (mainly time), this approach is often limited to organisms with short generation time that are also easy and cheap to keep, not to mention that their natural authenticity can come into question.

Alternatively, one can exploit natural "experiments" that have run for hundreds or thousands of years and with large, albeit less controlled, population sizes. A popular choice of such natural systems are *hybrid zones*.

Hybrid zones are areas where two parental types meet and mate, producing offspring of mixed ancestry (i.e. hybrids). Since we can still observe these contact zones several hundreds, thousands and millions of generations after they form, the maintenance of the distinct parental types is evidence for some kind of a *force* holding them apart despite dispersal. This force is well known in the field of evolutionary biology and it is called *natural selection*. The hybrid zone can be maintained by natural selection in several ways. For example, the hybrids may be less able to survive and to produce viable offspring, or there may be a barrier to interbreeding (or crossing) of the two parental types. Either way, hybrid zones became a valuable source of information on the strength of selection, as well as an argument in favour of its existence [4, 16, 57, 23].

The strength of selection acting on maintenance of the hybrid zone can be quantified by its "steepness", i.e. by numerically characterising how well the two parental species are kept apart. To study this, one typically follows the gradient for a trait typical for one of the parental

populations as one crosses the border from one parental population to the other. This gradient is also known as a *cline* and forms an sigmoid curve with two flat extremes inside the parental populations and a gradient in the area between them. The *centre of the cline* is then exactly at the point where the cline reaches the middle value (from the continuum of all possible values). The *width* of the cline can be defined as the distance between intersections of the tangent to the cline at its centre with the two horizontal lines at its extremes [4]. A hybrid zone may consist of several clines with shifted centres and different widths.

Hybrid zones with their component clines have been studied in many systems [4]. The most famous examples include tropical butterfly species of the genus *Heliconius* meeting in Latin America [33] and two toad species of the genus *Bombina* coming to contact around the Danube Basin and the Carpathians [65]. One popular way to describe a cline is by the frequency of any genetic difference that distinguishes the parental types (i.e. a SNP, indel or length polymorphism).However, clines can also be characterised by a changing phenotype. For example, a changing frequency of a binary feature such as presence of a patch on a butterfly wing, or by a continuously changing quantitative trait such as the average skin thickness of a toad [46], varying smoothly from one toad species to another.

In this thesis, we use a large dataset of genotypes and photographs coming from a hybrid zone between two subspecies of *Antirrhinum majus* with distinctly-coloured flowers. The clines in this hybrid zone are formed by frequencies of genotypes typical of the respective parental types, as well by frequency of flower colour phenotype.

1.3 Flower colour as a model for adaptive evolution

Colouration is highly conspicuous, accessible, studied and understood trait, which makes it a good model for both evolutionary genetics and understanding complex phenotypes [47]. To draw the link from colouration to adaptation, one needs to understand the genetic architecture of the colour trait and its effect on fitness.

A popular class of colouration models in genetic research is flower colour pattern. The tradition of studying flower colour systems is ancient and fruitful, which is not suprising given human captivation by flowers, which is provably at least 120,000 years old [29]. This attraction was eventually turned into commercial interest, manifesting in cultivation and flourishing trade with roses (from 5-th century BC, but especially from the 16-th century on) and tulips (from the 17-th century on). Quite early on, the gardeners noticed recurrent flower colour phenotypes and their proportions in offspring resulting from repeated crosses. However, it took until 1865 to establish the rules of heredity, when Gregor Mendel presented his work on pea and bean plants (genera *Pisum* and *Phaseolus*) at a meeting of Brno Natural History Society, featuring flower colour as a heritable trait and laying down the foundations of genetics [44].

Plants in general have also been popular models for studying hybrid zones, with 137 plant hybrid zones summarised in a recent review [1]. Quite understandably, in most of the plant hybrid zones discovered so far, the two parental types appear different to the human eye, many of them exhibiting differences in flower colouration.

One example is a contact zone of two distinctly coloured ecotypes of North American monkeyflowers *Mimulus aurantiacus*. The red ecotype typical for the coastal region of Western North America has been shown to be separated from the inland yellow ecotype by *ecogeographic isolation*, as the two ecotypes prefer distinct environmental conditions in nature.

Another type of isolation, associated especially with hybrid zones where the two parental types differ in floral traits is *ethological isolation* consisting of two behaviours: *pollinator preference* and *pollinator constancy* [21]. In the case of pollinator preference, the pollinator prefers visiting one type of flowers near others, for example in *Ipomopsis* and *Mimulus* hybrid zones where the two parental types differ in flower colour, morphology and nectar production attracting two distinct pollinators: humming birds and hawkmoths [2, 24]. On the other hand, in case of constancy the individual pollinator prefers to avoid switching from one flower type to another in consecutive visits, even if the species overall may have no visible preference for one or the other type. Thus, both pollinator preference and constancy results in assortative mating by reducing the transfer of pollen between the two parental types.

If the assortative mating promotes an advantage for the common type in the hybrid zone, pollinator behaviour may lead to formation and maintenance of stable clines. However, there may be a lot of gene flow everywhere in the genome except around flower colour genes found both in *Mimulus* [58, 61] and *Antirrhinum* [73] hybrid zones.

In our study system, the hybrid zone of two distinctly coloured subspecies of *Antirrhinum majus*, the assortative mating and frequency dependent selection seems to be facilitated by constancy of the bumblebee pollinators [66].

1.4 The *Antirrhinum majus* flower colour hybrid zone in Planoles

Antirrhinum majus is a relatively common plant in the Spanish and French Pyrenees [32], whose cultivated forms can be often found in gardens. It is also a well established plant system to study flower colour inheritance [71], flower development [12] and pollinator behaviour [64, 31, 45, 55].

Since the rediscovery of the hybrid zones between its two subspecies with differently coloured flowers in Spanish Pyrenees in the early 2000's (for more about the history of the hybrid zone, see [50]), *A. majus* has also been a model for studying adaptive evolution. More details about this hybrid zone can be found in [15].

In this thesis we study a hybrid zone of two subspecies of the common snapdragon *Antirrhinum majus*: *A. m. pseudomajus* with rich magenta coloured flowers and *A. m. striatum*, whose flowers are bright yellow. Although these two subspecies (also called "colour morphs" in some sources) meet at several areas across Spain and France, here we will focus on the hybrid zone at the Spanish side of the Pyrenees near the Catalan town of Planoles.

As is typical for *A. majus*, it grows both in patches and individually in disturbed soil, mostly along roads and on railway embankments, often out of cracks in the rocks and sometimes in blackberry bushes. The yellow subspecies can be found mostly in the west and the magenta one on the east side of the hybrid zone (figure 1.1) with an increased concentration of hybrids in the hybrid zone core.

The *Antirrhinum* hybrid zone near Planoles is particularly suitable for hybrid zone research due to its richness in hybrids and high population density, relatively long sampling history, good knowledge of the system and unprecedented sample size. In this thesis, we will use the dataset available as of 2019 with $22,358$ genotypes for 120-248 SNPs described in [3] from plants collected in years 2009-2019. These genotypes have been subject to several population genetic studies: besides, for example, calculations of inbreeding depression [3] and barriers to gene

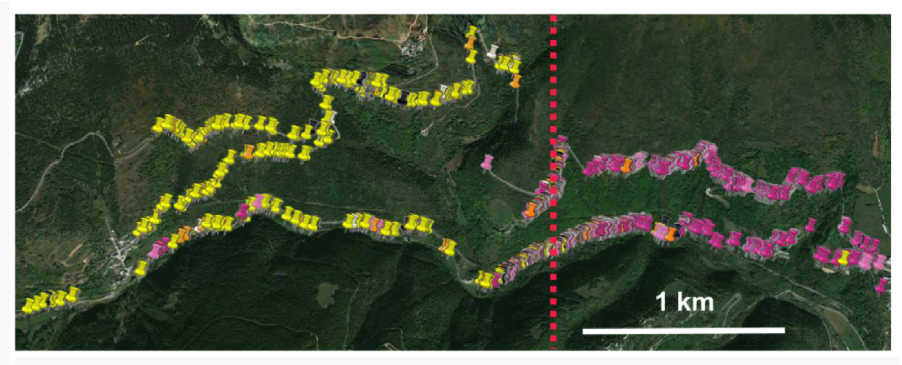


Figure 1.1: Map of the *A. majus* hybrid zone in Planoles with pins in different colours corresponding to six main flower colour phenotypes: yellow, weak orange, dark orange, white, pink and full magenta. The red dashed line shows the centre of the hybrid zone. Source: The group archive

flow [51], this SNP dataset is being used to build a pedigree consisting of individuals from the dataset (David Field, pers. comm.). The existence of these studies and resources, as well as interest of the community underlines the need for a systematic approach to measuring of the corresponding flower phenotypes.

The field data on flower phenotypes consist of flower photographs and manual colour scores: the *yellow*, the *red* (in fact magenta) and the *venation* score. All three scores depend on intensity, as well as on the distribution of the aurone and anthocyanin pigment. While red and yellow scores aim to describe the amount and distribution across the flower parts more generally, the venation score focuses on the intensity and spread of magenta colouration of veins on the inner side of the upper petal with higher values corresponding to more intense and wider spread venation pattern.

More specifically, the red score (originally used in [72], adapted and described in [67], figure 4.5) ranges from 0 to 5, in 0.5 intervals. In general, lower values of the red score correspond to flowers with less magenta (anthocyanin) pigment and vice versa. However, the score also reflects *distribution* of the pigment across the flower: scores of $1.5 - 2.5$ refer to magenta pigment present in isolated patches, whereas scores from 3 upwards represent flowers with smooth distribution of magenta pigment throughout the flower, with higher red score corresponding to flowers with more intense magenta colouration.

In the field, the table in figure 1.2 together with a verbal description of the crude categories (e.g. for red: $0 - 1$: pale, $1.5 - 2.5$: patchy and $3 - 5$: smooth) are used to assign the yellow and the red score. However, as the table suggest, deciding between values within the crude categories depends largely on the scorer and their experience. Furthermore, although the manual score is discrete by design (and makes sense, as it relies on human scorers), there is no evidence to support that the variability is, in fact, discrete. Rather the opposite: figure 3.5 of [72] shows that rather continuously varying amounts of anthocyanins found in flowers correlate with manual scores.

There have been at least two successful attempts to develop an objective and quantitative continuous colour measurement system for *A. majus* flowers from photographs (discussed in more detail in chapter 4). However, both of these methods required expert input on each flower photograph. This makes the methods impractical when dealing with large numbers of photographs and implies problems with reproducibility, as the human input is always, to

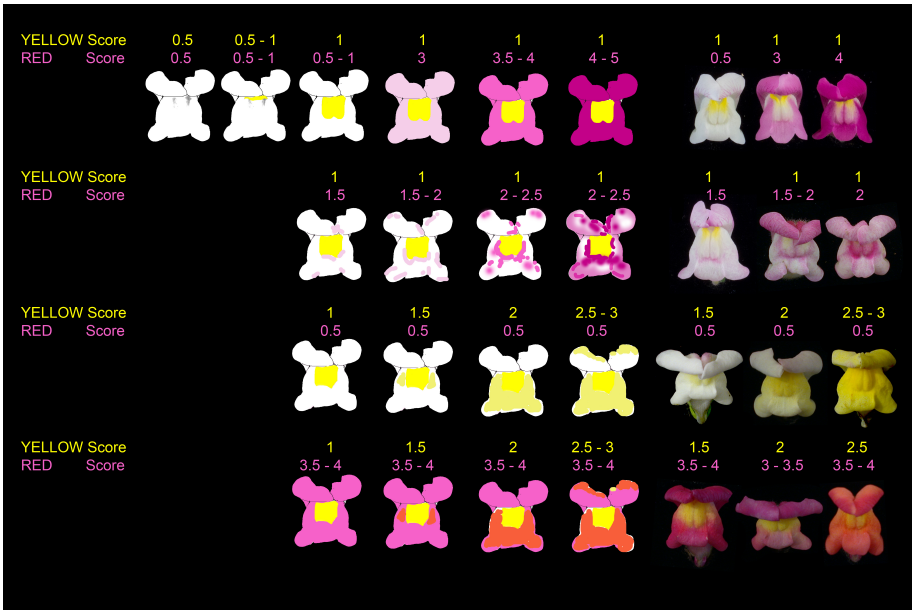


Figure 1.2: The flower colour scoring table used in the field. The scores depend on intensity, as well as on the distribution of the pigment. Scores are adapted from [72]. Image source: Group archive.

some extent, subjective. Since we are facing ten thousands of flower photographs with more to come, we strive to produce a fully automated pipeline to improve scalability and objectivity in addition to the requirements mentioned above.

1.4.1 Genetics of flower colour in the *A. majus* hybrid zone

Since wild *Antirrhinum majus* are self-incompatible [40] (i.e. cannot be fertilised by their own pollen), their fitness is dependent on bumblebee pollinators bringing in pollen from other potential mates. It has been shown that flower colour affects pollinator behaviour, especially search time and other foraging behaviours in bumblebee [60]. Furthermore, the standing hypothesis is that flower colour patterns, also called pollinator guides, help pollinators to enter the flowers and guide them through the floral tube of *A. majus* to the nectar and thus increase the pollinator's efficiency in transferring the pollen from one plant to another [7]. Therefore, the flower colour and its pattern are thought to have a significant effect on fitness of the *A. majus* plants.

The main pigments found in flowers of *A. majus* are yellow flavonoid *aurone* and magenta *anthocyanin*, whose expression is governed by a handful of loci [22, 63, 54, 73].

The dominant allele SULF of the *Sulfurea* locus, typical for *A. m. pseudomajus*, codes for an inverted duplication that generates small RNAs. These RNAs then repress the aurone biosynthesis gene, restricting the yellow pigmentation to a small yellow patch at the floral entrance [7]. On the other hand, the recessive plants including *A. m. striatum* individuals have a deletion in that area, leading to homogeneous aurone expression across the face and the tips of upper petals (the left column of flowers vs. the right in figure 1.3).

The regulation of anthocyanin expression, on the other hand, seems to be more complex. The two main players known so far are two tightly linked loci *Rosea* and *Eluta*. The *Rosea* locus consists of three MYB-like transcription factors with $\sim 90\%$ protein sequence identity referred

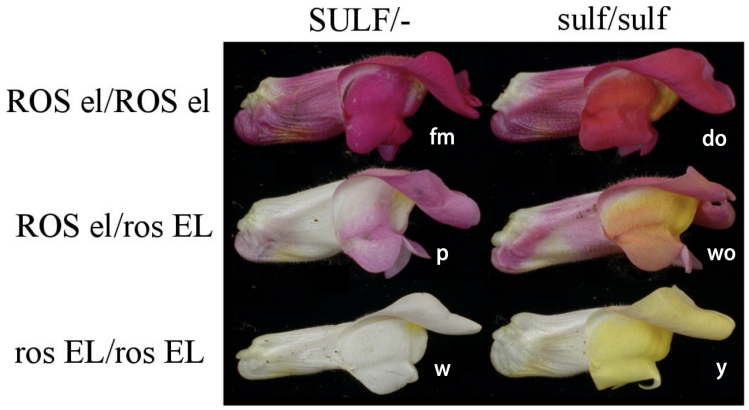


Figure 1.3: Typical phenotypes for the six Ros/El and Sulf combinations corresponding to the six main flower colour phenotypes: full magenta (fm) and dark orange (do), pink (p) and weak orange (wo), white (w) and yellow (y). Adapted from [15].

to as ROS1, ROS2 and ROS3, respectively [68]. It has been shown that ROS1 and ROS2 bind to DNA, promoting expression of the anthocyanin biosynthetic gene, with ROS1 explaining most of, and ROS2 a small portion of, the variation in anthocyanin expression [54]. Therefore, when talking about the *Rosea* locus, we will mostly refer to ROS1. It has also been shown that there is an enhancer of ROS1 downstream from ROS1, explaining most variation in ROS1 expression [68, 67].

Another relevant locus in anthocyanin expression is *Venosa*, whose dominant (VEN) allele typical for *A. m. striatum* allows for expression of anthocyanin in veins on the inner side of upper petals [54].

The *Eluta* locus, on the other hand, restricts anthocyanin into central areas of the face and hence, the flowers carrying a copy of the dominant *Eluta* (EL) allele have all the anthocyanin restrained into patches in the central part of the face of the flower and the venation into a small patch on the centre of the inner side of the upper petal (i.e. at the area of the bee entrance [68]).

The two typical *Rosea-Eluta* (Ros/El) genotypic combinations are rosEL for *A. m. striatum*, mostly west, and ROSel for *A. m. pseudomajus*, mostly east of the hybrid zone core. There is only very little anthocyanin in flowers of homozygous rosEL plants, classifying them as white or yellow (depending on *Sulfurea* locus) out of the six canonical phenotypes in figure 1.3, with typical manual red score 0.5 (figure 1.2). On the other hand, in flowers of homozygous ROSel plants the anthocyanin is produced quite homogeneously throughout the corolla, with a just one pale patch at the flower entrance, making them fall into the full magenta or dark orange category (figure 1.3), with manual red scores of 3 and above (figure 1.2).

The results from crosses in section 4.2.2.1 of [67] showed that the two loci, *Rosea* and *Eluta* lie 0.5 cM apart. The evidence for selection has been argued in [73] and [68]. Firstly, the presence of steep clines requires selection unless there has been very recent secondary contact. Ruling that out is not completely trivial, but the proportion of recombinants suggests that the hybrid zone is at least 100 generations old (SI 9 in [68]). Secondly, positions of *Rosea* and *Eluta* coincide with two sharp F_{st} peaks driven by a drop in diversity within the two parental

populations, implying selective sweeps at these two loci. Combining some older estimates of dispersal ($645m(\pm 40m)$) from a large-scale pedigree in the hybrid zone with the *Rosea* cline width estimates ($2000m$) imply moderately strong selection with selection coefficient of about $0.42(\pm 0.05)$ (all estimates and calculations from section 6.2 in [15]). Although the exact mechanism of this selection is not clear, preliminary pedigree data show that the rare, hybrid phenotypes have consistently lower fitness and there also seems to be frequency dependent selection favouring the common type (David Field, pers. comm.)

From what we already know, the effects of the *Ros/El* locus on anthocyanin expression in *A. majus* flowers are quite complex and investigating them further is challenging, especially since *Ros* and *El* are so closely linked. To address these issues in this thesis, we will use the advantages of our dataset. Firstly, the rich variety of diverse phenotypes and genotypes (including *Ros/El* recombinants) coming from the natural conditions of the hybrid zone can help us explore the entire genotype to phenotype relationship in its complexity and put us in a more comfortable position when talking about "the whole picture". Secondly, the presence of flower images (as opposed to manual scores alone) for an absolute majority of sampled plants allows us to go into great detail in studying the phenotype. Last but by no means least, the sheer numbers of the sampled plants provide us with unprecedented statistical power to disentangle the effects of closely linked loci. However, to exploit all the advantages this dataset has to offer, we need to use scalable and automatic methods capable of dealing with such large datasets, which we develop, introduce and, hopefully, put in a good use in the following chapters.

1.4.2 The null model

It has been established that the majority of variation in anthocyanin pigmentation in *A. majus* flowers can be attributed to differences in ROS1 and EL [54]. However, we already know that there is more to it. Apart from *Venosa* causing venation, there is an enhancer downstream of ROS1 modulating the expression of ROS1, several alleles of ROS are described in commercial stocks of *A. majus* and the exact location of recombination between ROS and EL all seem to have an effect on anthocyanin production and distribution in flowers of *A. majus* [67, 68].

Similarly, since ROS and EL are so closely linked, they are usually inherited together forming two predominant Ros/El genotypes *mostly* corresponding to two distinct subspecies, their sharp clines defining the centre of the hybrid zone. However, that is also not the whole story. We already know that the anti-correlation between ROS and EL is not absolute due to recombination and there are various asymmetries regarding the Ros/El recombinants, as well as dissimilarities in the shapes of *A. m. pseudomajus* and *A. m. striatum* clines. For example, there seems to be fewer *rosel* than *ROSEL* recombinant haplotypes found in the hybrid zone (1.6% vs. 3.3% in table 6.2 in [67]), although theoretically, they should be generated at an equal rate. Also, there seems to be more than twice as many *A. m. pseudomajus* haplotypes on the west from the centre of the hybrid zone (i.e. on the "*A. m. striatum*" side 6.9%) than there is *A. m. striatum* haplotypes on the east side from the centre of the hybrid zone (i.e. on the "*A. m. pseudomajus*" side, 3%, section 6.2.2 in [67]).

To distinguish the expected Ros/El genotype patterns and the major, relatively well-known flower colour pattern effects of ROS1 and EL from those less studied, we developed a simplified two-locus Ros/El *"null model"* that describes the hybrid zone between *A. m. pseudomajus* and *A. m. striatum* and how these two major loci affect the expression and distribution of the magenta pigment in flowers (both based on [68]).

The two-locus Ros/El model

Rosea (Ros) and *Eluta* (El) are *two* closely linked *loci*, with *two alleles* each. For both of these loci, selection favours one allele at one side of the hybrid zone and the second allele at the other side. Thus, selection is maintaining steep clines at both of the loci in a balance against dispersal. The amount of anthocyanin in flowers is governed by these two loci, the dominant *ROS* allele coding for high production of anthocyanin, the semidominant allele *EL* restricting any anthocyanin produced by *ROS* to a few centrally located patches. In this *null-model*, we define *ROS* as synonymous with *ROS1*.

From now on, we will use this model throughout the thesis to put our findings into perspective and to identify discrepancies from what is expected.

1.5 Motivation

To study evolutionary questions using data from hybrid zones, one needs to understand the genotype-phenotype-fitness map. Using objective and scalable methods on a large dataset helps us to get a better grip both on the genotype and on the phenotype, but most importantly, on their relationship.

The scope of data from the *Antirrhinum majus* hybrid zone in Planoles is unprecedented. There is a great potential to measure several parameters defining the system with unusual accuracy, but to really profit from this rich dataset, we have to exploit as much of it as we can. That is why we introduce several ways of dealing with large datasets and their applications to the dataset in question.

First, we need to get a good understanding of the genotype. The genotypes we measure do not, by default, tell us which alleles are inherited together from the same parent. In other words: from genotypes we do not know the *haplotype*. Haplotypes tell us which alleles are on the same molecule of DNA, i.e. which alleles get transcribed and translated together, so this information may be relevant to study the relationship of genotype to phenotype and anything to do with heritability. Therefore, in chapter 2 we use the advantage of large sample size to estimate the partial haplotypes in the *Ros/El* region for each genotyped plant, using the Expectation Maximisation (EM) algorithm. In the same chapter, we also investigate haplotypic structure of the wild population in a locus that is under selection and forms a stable, steep cline, while explaining the majority of variation in flower anthocyanin pigmentation. We study association of the haplotypes with flower colour phenotypes, relationships between the haplotypes and their geographical distribution, striving to answer questions like: Are the similar haplotypes associated with similar flower colour phenotypes and do they inhabit the same geographical areas? Is the extent of variation in the *Ros/El* region the same for both subspecies? Which are the most common haplotypes?

In chapter 3 we use the inferred haplotypes together with the flower scores to extend our knowledge on plants carrying the recombinant *Ros/El* haplotypes, including the effect of rare genotypes on the flower colour phenotype. More specifically, we use the large numbers of plants sampled in the wild to investigate the effects of various recombinant *Ros/El* genotypes on the flower colour phenotype and compare the frequencies of and the variation within the two recombinant and the two parental groups of *Ros/El* haplotypes to what is expected. In this chapter we discover discrepancies to phenotypes expected given the genotype (in particular, we identify several haplotypes that defy what is expected), diagnose unexpected missing variation

in *A. m. pseudomajus* haplotypes and confirm the curious deficit of *rosel* recombinants in the hybrid zone.

To improve flower colour information from the discrete manual scoring system, we developed an automated consistent, flexible and scalable system for detecting flower colour and its patterns. In chapter 4 we present our pipeline to clean the flower images from the background, normalise their lighting conditions and finally, to output an array of colour measurements from each photograph. There, we also test it, compare the manual scores to the machine-measured scores, show that photographic measurements correlate with amounts of pigments measured using spectrophotometry and investigate the effect of different lighting and setups on the machine scores.

Finally, in chapter 5 we use the machine-measured flower colour scores to fit and examine a phenotype cline across the hybrid zone in Planoles using full flower images as opposed to discrete, manual scores. We then compared clines formed by several measures derived from the machine flower colour scores. We showed that the phenotypic clines defined by the machine colour scores are concordant with the genotypic clines.

CHAPTER 2

Haplotypic variation in the hybrid zone

To understand the genotype-phenotype-fitness map we need to understand the genotype first. In particular, we are interested in the space of possible genotypes available in the wild, how they are distributed across the geographical region and how they relate to phenotypes. Are there any constraints on the genotypes, or are all possible genotypes equally represented in the population? Can we observe a pattern in the association of genotypes to phenotypes?

To better understand the structure of genotypes in the population, we reduce the problem to simpler units: to *haplotypes*. As opposed to genotypes, which are just a summary of genetic information at each genetic locus, haplotype carries the information about the physical molecule of DNA the genetic marker sits on, i.e. the chromosome. This means, that rather than a set of summaries of markers at each genetic locus, with haplotypes we know which markers are inherited together on the same chromosome, coming from the same parent. This is important, because it helps us to understand the inheritance and genetic patterns circulating in the population. Since our model system is diploid (an individual carries two copies of every genetic region), we will need to find out how do the genotypes "break" into two haplotypes for each individual (see figure 2.1).

In this chapter we focus on the genetic region near *Rosea* and *Eluta*, the two closely linked loci that have been shown to explain majority of variation in anthocyanin expression in flowers of *Antirrhinum majus*. We estimate the most likely *partial* haplotypes in this genetic region and their frequencies in a wild hybrid population of *A. majus* using the Expectation Maximisation algorithm (EM). We explain our implementation of EM together with its extensions and we analyse the resulting phasing and haplotypic frequencies in a sample of more than $22,000$ plants. Especially, we study the structural similarity among haplotypes together with their abundance and we link them to flower colour phenotypes and geographical positions of plants carrying them.

In chapter 3, which is an extension of this chapter, we study the structure of the haplotypes in relation to *Rosea* and *Eluta*, the two closely linked loci that have been shown to explain the majority of variation in anthocyanin expression in flowers of *A. majus*. In particular, we study haplotype structure in the context of the two parental and two basic recombinant types of *A. majus* and correlations between the individual markers based on their positions relative to *Rosea* and *Eluta*.

There are several experimental and computational methods used for haplotype phasing, reviewed in [9]. Since our dataset consists of more than $20,000$ individuals, experimental approaches

are out of the question and we have to stick with computational approaches. Since *A. majus* is not a very widely-used study system, we do not have an accurate prior knowledge of haplotype frequencies in the population and the available pedigree only includes a small fraction of the individuals. Considering also the relative sparsity of available SNP genotypes ($120 - 240$ markers per plant genome), in this pilot project we decided to focus on the partial haplotypes centered around the crucial genes *Rosea* and *Eluta*, with $12 - 24$ markers in 1 cM. We settled on using the Expectation Maximisation algorithm (EM) [14], which is sufficient and well-suited for a problem of this scale. While PHASE [62] might be a popular choice for similar problems, it is designed for smaller sample sizes (of "up to several hundred individuals" according to [9]). Furthermore, other tools are often limited in number of unique local haplotype states and they estimate the phase locally. This is crucial when dealing with longer genotypes. However, this could lead to omission of rare haplotypes and also to bias when estimating linkage from the results (and in our case this is also unnecessary). On the other hand, EM is simple, easily interpretable, considers all possible 2^{12} haplotypes and most importantly, it does not make use of information on genetic proximity, so we will be able to make statements about linkage disequilibrium.

2.1 From genotypes to haplotypes using the Expectation Maximisation algorithm (EM)

2.1.1 Genotypes and haplotypes

Haplotypes are represented by strings of 0s and 1s, each digit representing a genotyped locus occupied by common and rare alleles respectively. A **genotype** of a diploid individual is represented by a string of 0s, 1s, 2s and 9s, each digit representing one genotyped locus (figure 2.1). 0s and 2s denote loci for which the individual is homozygous, 1s stand for heterozygous loci and 9s for loci that have not been genotyped. Please note that in real data, missing values are denoted with -9 (failed to genotype) and -10 (purposefully not genotyped in the given plant), but here we use 9 in both cases for simplicity. The fraction of successfully genotyped loci differs from plant to plant. After excluding the genotypes of poor quality, there are less than 1% loci with missing genotypes left in the dataset we use in this chapter.

2s typically correspond to loci homozygous for the "rare" allele. Assignment of an allele as "rare" is arbitrary and depends on the dataset. In our dataset, the 2s were chosen such that they are historically more common in the magenta *A. m. pseudomajus*, whereas 0s tend to be associated with yellow *A. m. striatum*. This is, however, not necessarily the case in our sample. Although different notations exist where 1s denote loci homozygous in rare allele and 2s stand for heterozygous loci, here we stick to this, more algorithm-friendly alternative, where one only needs to "add" the two haplotypes together locus-wise to obtain the diploid genotype.

The SNP dataset used here consists of $22,358$ LGC genotypes on 120 - 240 SNPs described in [3]. Since the running time and memory required by the expectation maximisation algorithm (EM) are in principle exponential in the number of loci, it would not be possible to calculate the haplotypes for the entire genome (or even entire linkage groups) this way. Furthermore, even with more than $22,000$ genotypes, there would still be too few individuals *per* unique genotype, making the estimates of haplotype frequency inaccurate. However, since we are mainly interested in anthocyanin pigmentation of the flowers, this is not necessary and we can focus on a smaller subset of 24 SNPs within 200 kb of the Ros/El region, summarised in table 2.1, and visualised in figure 3.8.

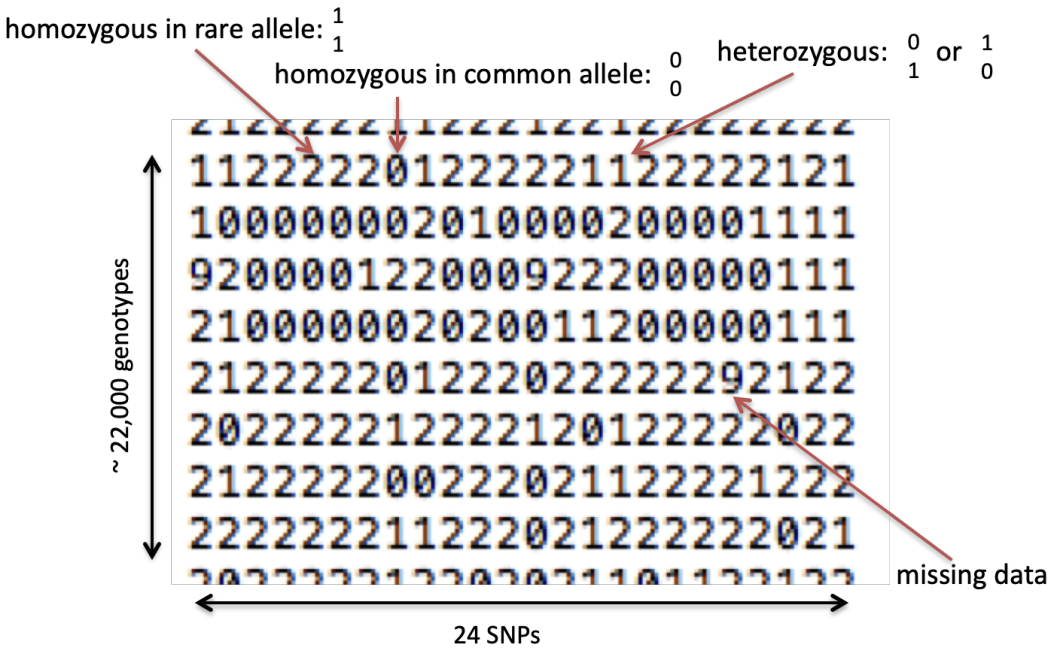


Figure 2.1: An excerpt of genotypes in our dataset with examples of loci and their interpretations into haplotypes. A genotype of an individual is represented by a string of 0s, 1s, 2s and 9s, each digit representing one genotyped locus. 0s and 2s denote loci for which the individual is homozygous, 1s stand for heterozygous loci and 9s for loci that have not been genotyped*. 2s typically correspond to loci homozygous for the "rare" allele. While the interpretation of 0s and 2s from genotypes into two haplotypes is unambiguous, 1s are a challenge, because one does not know on which chromosome there is a 1 and on which chromosome there is a 0. This is particularly problematic, if there are several 1s in a genotype, as these could have originated from several different haplotype combinations. For example, a a genotype ...1...1... can be broken into haplotypes ...0...1... and ...1...0... or into haplotypes ...0...0... and ...1...1.... * Please note that in real data, missing values are denoted with -9 (failed to genotype) and -10 (purposefully not genotyped in the given plant).

Since fewer than 30% plants out of $22,358$ were genotyped at all 24 SNPs in the Ros/El region, we decided to consider two datasets: the *"numerous"* dataset on 12 SNPs genotyped in all plants and the *"long"* dataset on all 24 SNPs, with about $6,500$ available plant genotypes. This way, we can exploit both the statistical power coming from the sheer number of individuals in the *numerous* dataset and the finer information coming from *long* dataset, where the individuals were genotyped in more detail.

All algorithms and calculations will be discussed here, the results from 12-marker *numerous* dataset will be discussed here and in chapter 3, while the results from the *long* dataset will only be discussed in chapter 3, especially in 3.3.

2.1.2 The EM algorithm

The EM algorithm [14] is an iterative approach to maximising likelihood for statistical models where the equations are not directly solvable due to a combination of latent variables and unknown parameters. Based on guarantees stemming from numerical methods, the algorithm cycles through series of expectation (E) and likelihood-maximisation (M) steps, increasing

id	LocusName ros_assembly_	V2_pos chr6	Type	Nearby genes	id in dataset *numerous*
0	473914	52249483	Diagn.	Am06g36380 intron	0 (left)
1	490488	52266149	Diagn.	Am06g36390 - Am06g36400	-
2	541834	52318288	ROS1	ROS1 exon1	-
3	543443	52319882	ROS1	ROS1 intron2	1 (Ros)
4	544601	52321040	ROS1	ROS1 exon3	-
5	567004	52343442	ROS2	ROS2 exon3	-
6	575837	52352415	ROS3	ROS3 exon3 (T2)	2 (Ros)
7	576271	52352849	ROS3	ROS3 exon3	3 (Ros)
8	618376	52395000	Diagn.	Am06g36520* exon3	-
9	620992	52397616	Diagn.	Am06g36520 - Am06g36530	4 (between)
10	635819	52412525	Diagn.	Am06g36560 - Am06g36570	-
11	653015	52429918	Diagn.	Am06g36580 - Am06g36590	5 (between)
12	660344	52437098	Diagn.	Am06g36580 - Am06g36590	-
13	670530	52447357	Diagn.	Am06g36610 exon6 (T2, T3)	6 (between)
14	674756	52451583	Diagn.	Am06g36610 intron	-
15	689955	52466782	Diagn.	Am06g36650 - Am06g36660	-
16	702909	52479781	Diagn.	Am06g36670 - Am06g36680	-
17	715001	52491887	EL	EL exon3	7 (El)
18	715015	52491901	EL	EL exon3	-
19	717957	52494503	Diagn.	indel region	-
20	737420	52514037	Diagn.	Am06g36720 - Am06g36730	8 (right)
21	744403	52521020	Diagn.	Am06g36750 exon3	9 (right)
22	748981	52525598	Diagn.	Am06g36780 - Am06g36790	10 (right)
23	758578	52535195	Diagn.	Am06g36780 - Am06g36790	11 (right)

Table 2.1: All genotyped SNPs available in our dataset from the region near *Rosea* and *Eluta*, the two closely linked genes playing a key role in production and distribution of magenta pigment in flowers. The six-digit code *xxxxxx* can be used to construct the names of these SNPs in the form *ros_assembly_xxxxxx*. The type is either a *Rosea* or *Eluta* gene the SNP is linked to, or "Diagnostic", if it lies outside of the two focal genes. The position on chromosome 6 in base pairs and position relative to the nearest genes is according to *A. majus* version 2 genome published in [40]. The position relative to genes is given to the nearest exon/intron with a note on transcript if it is not present in all transcripts (*Tx* if it is only transcribed in transcript *x*) and it is noted as *X* - *Y*, if the SNP is located between gene *X* and gene *Y*. The presence of the SNP in the *numerous* dataset of 12 loci, their id therein and a position category (in parentheses) used in linkage disequilibrium analysis (section 3.2.2) are indicated in the last column. The line separating first twelve SNPs from the second twelve SNPs denotes the border between *long1* ("around Ros") and *long2* ("around El") SNPs as referred to in section 3.3.

the likelihood in each step and converging to a set of parameters corresponding to a (local) likelihood maximum (for more details see [14]).

The use of the EM algorithm in this scenario was inspired by [30], expanded to use missing data and to allow for prior information. The algorithm takes a list of genotypes and their counts in the sample ($\mathbf{n_g}$), a list of possible haplotypes, and a vector of initial frequencies of these haplotypes $\mathbf{p}^0$ as an input. It then iterates over the estimated frequencies $\mathbf{p^t}$ via a cycle of expectation and maximisation steps, until the change between two subsequent iterations of $\mathbf{p}$ is lower than a defined threshold ε.

Algorithm 2.1: EM Algorithm for haplotype phasing

Result: $\mathbf{p^t}$

$\mathbf{p^t} = \mathbf{p^0}$;

$\delta = \infty$;

while $\delta \geq \varepsilon$ **do**

  calculate the expectations $P(h_j, h_k \mid g_i)\ \forall i, j, k$;

  maximise likelihood: update $\mathbf{p^{t+1}}$;

  $\delta = ||\mathbf{p^{t+1}} - \mathbf{p^t}||$;

end

Summary of symbols used in the EM algorithm

$\mathbf{n_g}$: the vector of genotype counts, where n_{g_i} is the number of individuals with genotype i in the sample

n: the number of all genotypes in the sample such that $n = \sum_i n_{g_i}$

$\mathbf{n_h^t}$: the vector of estimated haplotype counts after iteration t, where n_{h_j} is the estimated number of haplotypes j in the sample, such that $\sum_j n_{h_j} = 2n$

$\mathbf{p^t}$: the vector of estimated haplotype frequencies after iteration t, where p_j^t is an estimated frequency of haplotype j after iteration t

$P(h_j, h_k \mid g_i, \mathbf{p^t})\forall j$: probability that genotype g_i is a result of combining haplotypes h_j and h_k given the estimated haplotype frequencies $\mathbf{p^t}$

S_i: a set of all pairs $\{j, k\}$ such that a combination of haplotypes h_j, h_k forms genotype g_i. The size of S_i is 2^{het_i}, where het_i is the number of heterozygous loci in genotype g_i.

Without missing data

In the **expectation step** we calculate the probabilities that a given genotype was formed by all possible combinations of haplotypes. If a genotype is homozygous in all loci, this is trivial. However, if a genotype can be formed by a combination, or several combinations of two different haplotypes (S_i), the probability is given by a conditional probability

$$P(h_j, h_k \mid g_i)^{t+1} = \frac{P(h_j, h_k)}{P(g_i)} = \frac{2p_j^t p_k^t}{\sum\limits_{\{j',k'\}\in S_i} 2p_{j'}^t p_{k'}^t},$$

for all $j \neq k$. This probability is either 0 or 1 if $j = k$, so that all loci are homozygous.

Maximisation step The role of the maximisation step is to find $\mathbf{p^{t+1}}$ which maximises the likelihood of the data given $\mathbf{p^t}$. The most likely numbers of haplotypes in each step can be calculated as

$$n_{h_j}^{t+1} = \sum_i n_{g_i} \alpha_{ji}^{t+1},$$

where α_{ji}^{t+1} is a contribution of count of genotype i to the estimated number of haplotype j.

$$\alpha_{ji}^{t+1} = \begin{cases} 0, & \forall i\text{: } \{j,k\} \notin S_i \text{ for any } k \\ 2, & \forall i\text{: } S_i = \{\{j,j\}\} \\ P(h_j, h_k \mid g_i)^{t+1}, & \forall i\text{: } \{j,k\} \in S_i,\ j \neq k \end{cases}$$

The contribution α_{ji}^t is a product of $P(h_j, h_k \mid g_i)^{t+1}$ and a number of h_j in $\{h_j, h_k\}$, i.e. 2 for any t, if g_i consists of two haplotypes h_j and it is 0 for any t if there is no h_k such that h_k and h_j form g_i together. In fact, the matrix $\mathbf{A^t} = [\alpha_{ji}^t]$ in this case is very sparse, which we use to speed up the calculations.

In the end, the most likely $\mathbf{p^{t+1}}$ can be calculated simply as their estimated proportion in the population of all haplotypes

$$\mathbf{p^{t+1}} = \frac{\mathbf{n_h^{t+1}}}{2n}$$

.

With missing data

Sometimes, there is a missing genotype at some SNPs. In the real dataset, it is denoted with a "-10" (not genotyped), or "-9" (genotyping failed), but in reality, it could have been 0, 1, or 2 and therefore all of these possibilities have to be included. Typically, this means that genotype with a missing value can be a result of more combinations of haplotypes, i.e. the contribution matrix $\mathbf{A^t}$ is less sparse. Furthermore, a haplotype can now be combined with more haplotypes to form a genotype with missing value, which makes some elements of $\mathbf{A^t}$ more complex. Adding insult to injury, the genotypes with missing data have to be separate from the complete genotypes, which means there would possibly be more genotypes in total, which would make the contribution matrix $\mathbf{A^t}$ larger.

The **expectation step** is similar to the one before, with more complex $P(g_i)$ and a possibility that $j = k$:

$$P(h_j, h_k \mid g_i)^{t+1} = \frac{P(h_j, h_k)}{P(g_i)}$$
$$= \begin{cases} \dfrac{2p_j^t p_k^t}{\sum\limits_{\{j',k'\} \in S_i, j' \neq k'} 2p_{j'}^t p_{k'}^t + \sum\limits_{\{j',j'\} \in S_i} (p_{j'}^t)^2}, & \forall \{j,k\} \in S_i \text{ for any } j \neq k \\ \dfrac{(p_j^t)^2}{\sum\limits_{\{j',k'\} \in S_i, j' \neq k'} 2p_{j'}^t p_{k'}^t + \sum\limits_{\{j',j'\} \in S_i} (p_{j'}^t)^2}, & \forall j : \{j,j\} \in S_i \end{cases}$$

In the **maximisation step**, the contributions to a given haplotype from a given genotype can now come from several combinations of haplotypes:

$$\alpha_{ji}^{t+1} = \sum_{\{j,k\}\in S_i, j\neq k} P(h_j, h_k \mid g_i)^{t+1} + 2\eta_{j,j|i} P(h_j, h_j \mid g_i)^{t+1},$$

where

$$\eta_{j,j|i} = \begin{cases} 1, & \text{if } \{j,j\} \in S_i \\ 0 & \text{otherwise.} \end{cases}$$

Since allowing for missing data would result in more possible genotypes, thus enlarging the contribution matrix $\mathbf{A^t}$ and making the contribution more complex by forcing us to allow for more possible combinations of haplotypes at the unknown sites, it leads to a significant increase in computation time. The data with missing genotypes can still be used and should be used (especially when the same genotypes with missing data are found repeatedly), but filtering the genotypes based on number of missing loci is a reasonable thing to do. For example, we excluded all the genotypes on 12 loci with more than 3 missing values in most runs of the EM in this thesis, as high number of missing values might indicate a failed genotyping attempt, genetic abnormalities, or poor quality of the individual genotyped sample.

With prior information

Analysing the *long* dataset in an attempt to exploit more densely genotyped individuals leads to a new kind of challenges. There are 2^{12} times more possible haplotypes on 24 loci (namely 2^{24}) than there are on 12 loci (2^{12}) and the sample size is only about the tenth of the sample size in *numerous* dataset. This means, that on average each haplotype is presented in much lower numbers and the frequency estimates will not be as accurate as in the previous case. However, 12 out of these 24 loci have already been analysed in the *numerous* dataset and therefore, this information can be take into account.

The difference in comparison to the previous versions of the EM algorithm is to replace the calculation of the updated $\mathbf{p^{t+1}} = \frac{\mathbf{n_h^{t+1}}}{2n}$ in the maximisation step with one that takes into account previously estimated frequencies $\mathbf{p}^*$ of *relevant* partial haplotypes. A partial haplotype (for example a haplotype from the *numerous* dataset on 12 loci) is *relevant* to a full length haplotype (for example from the *long* dataset), if the partial and full length haplotypes are equal at all loci genotyped in partial haplotype (while all loci genotyped in partial haplotype are also genotyped in the full length haplotype). For example, if a partial haplotype "11" is genotyped at loci 1 and 3 of a full length haplotype on three loci, we can also represent it as "1.1" and it is relevant to full length haplotypes "101" and "111", but not to "000", "001, "011", etc. This means, that a partial haplotype on l loci is relevant to 2^{m-l} full length haplotypes on m loci. Let us denote R_j the set of all of full length haplotypes to which partial haplotype h_j^* is *relevant*. Then we can estimate the elements of $\mathbf{p^{t+1}}$ as fractions of frequencies of their *relevant* partial haplotypes distributed according to the estimated numbers $n_{h_k}^{t+1}$ of all the full length haplotypes h_k from R_j:

$$p_k^{t+1} = p_j^* \frac{n_{h_k}^{t+1}}{\sum_{k\in R_j} n_{h_k}^{t+1}}$$

Since $\mathbf{p^{t+1}}$ is used in the next round of calculating $P(h_{k_1}, h_{k_2} \mid g_i)^{t+2}$, the information on the partial haplotypes will stay in the system throughout the whole cycle of the EM algorithm. This is helpful, since it is making sure we are not attempting to estimate frequencies of partial

haplotypes again, from a much poorer dataset (remember, that the dataset of full length haplotypes has much lower sample size). At the same time, it is not excessively affecting the real estimates of frequencies of full length haplotypes *in the sample* that will result from the actual numbers of genotypes $\mathbf{n_g}$ in the sample *and* the estimated contribution matrix $\mathbf{A^t}$ resulting from the EM. It is quite important to note that using information on the partial haplotypes when estimating the full length haplotype frequencies only makes sense when both of the datasets come from the same population and when the partial haplotype dataset is much more plentiful. Both of these requirements are met in our case, so we use this approach on the *long* dataset, using the partial haplotype frequencies estimated previously from the *numerous* dataset.

This extension works well together with the extension for missing data and does not significantly affect the running time or memory requirements.

2.1.3 Other measures

For each haplotype h_j, the estimated *support* of the haplotype n_{h_j} is a sum of numbers of observed genotypes n_{g_i} such that $j \in S_i$ (possibly formed by that haplotype) multiplied by the proportions that the genotype contributed to this haplotype α_{ji}:

$$n_{h_j} = \sum_i n_{g_i} \alpha_{ji} = [\mathbf{p} \circ \mathbf{n_g}]_j$$

where α_{ji} and $\mathbf{p}$ are the output of the EM algorithm (and $\circ$ is a symbol of element-wise multiplication). In other words, *support* of a haplotype is simply a point estimate of the number of copies of this haplotype in the sample. Since our sample consists of n diploid genotypes, the number of copies of all haplotypes combined $\sum_{j \in S_i} \alpha_{ji} = 2$ and $\sum_j n_{h_j} = \mathbf{p} \cdot \mathbf{n_g} = 2n$.

This allows us to calculate point estimates of **mean red score** associated with each haplotype as a weighted mean of the manual red score (described in refsec:intro-dataset) over all plants k that could have had the haplotype:

$$\bar{r}_j = \frac{\sum_{k \in \{1,2,\ldots n\}} \alpha_{ji} r_k}{\sum_{k \in \{1,2,\ldots n\}} \alpha_{ji}}$$

for every plant k with available red score r_k, where i is the genotype of plant k. Note that if red scores were available for all n plants, the sum in denominator would be equal to n_{h_j} for every haplotype j.

Similarly, we can calculate point estimates for **mean latitude and longitude**, or **mean northing and easting**, respectively.

2.2 Results

The estimated haplotype supports (see figure 2.2), i.e. estimated real number of copies of that haplotype in the sample of $22,358$ plants range from $3e^{-61}$ (most likely not present at all) to about $5,588$ copies present in the population. This maximum support corresponds to the presence of the most frequent haplotype in about $4,576$ plants (506 of these are homozygous).

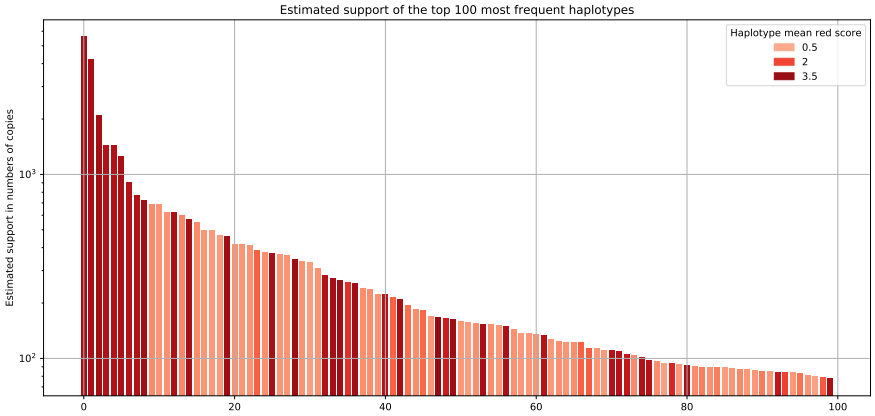


Figure 2.2: The estimated support of the top one hundred of the *numerous* haplotypes in shades of red representing its estimated mean red score. There seem to be fewer dark red haplotypes, present in higher numbers, while there seem to be many more less common, more variable haplotypes associated with lower red scores.

Interestingly, the dark red haplotypes seem to be overrepresented on the left side (more common haplotypes), while the haplotypes associated with lower red scores seem to be much more variable, although present in lower numbers.

To avoid misinterpretation of the haplotype mean red score $\mathbf{r}$, one should add that due to dominance, the mean red score will depend on the genetic background as well. Let us remember (see 1.4.2), that *ROS* is dominant over *ros* causing higher amounts of magenta pigment in the flowers. On the other hand, *EL* is dominant over *el*, causing restriction of the present magenta pigment into centralised patches, typically resulting in intermediate red score values of $2 - 2.5$ in individuals carrying both *ROS* and *EL* alleles. Firstly, this means that the more the haplotype is present in plants in the east side of the hybrid zone, the higher is the chance that it will be paired with dominant *ROS* haplotypes prevalent on the eastern side and hence, the higher its mean red score will be no matter the *Rosea* allele it carries. On the other hand, *ROSel* haplotypes that would typically induce a fully magenta coloration with red score $3.5 - 4.5$ in homozygotes will be restricted by the typical western *EL* allele to red scores in range $2 - 2.5$ instead of $3.5 - 4.5$ that would be typical for it in its homozygous state. Secondly, the smaller number of the plants the haplotype is indicated in and the more central its mean geographical position, the less sure we can be of its true red score (especially if the estimated mean red score is intermediate).

As expected, the mean haplotype geographical positions are less extreme and more central than those of the real plants (figure 2.3). Partly, this is caused by the fact that 57% of the genotyped plants were collected from within one kilometre in the hybrid zone *core*. Also, the location of the estimated mean red score corresponds to the real positioning of the hybrid zone.

In the principal component (PC) space created by the space of all haplotypes present in numbers of times corresponding to their support ("fortified haplotype space", figure 2.4) we can see that the first principal component separates the dark red haplotypes from the paler ones. Since all 2^{12} haplotypes are present in the dataset, the data has a shape of a twelve-dimensional hypercube. This leads to spurious structure of the data projected into components of PCA. This structure cannot be removed by using weighted PCA, which would only stretch the hypercube in heavily weighted directions, nor can it be removed by choosing

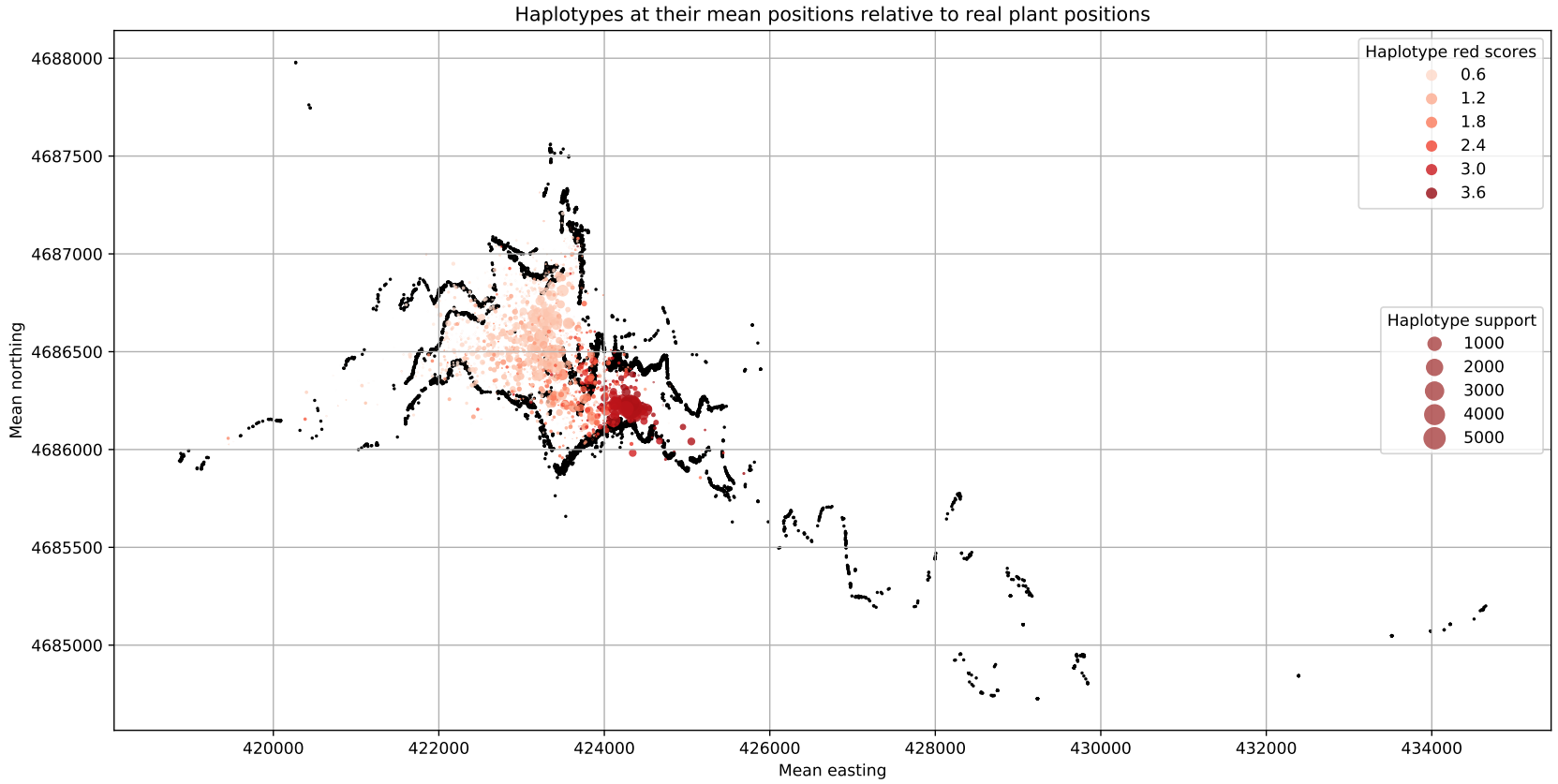


Figure 2.3: The haplotypes at their estimated mean geographical positions (discs in shades of red) plotted on the background of all plants in the dataset at their real positions (in black). The shades of red denote the estimated mean red score associated with each haplotype and its support in the data is represented by the size of the marker. The dark red haplotypes seem to be more concentrated, larger and less variable than the pale haplotypes.

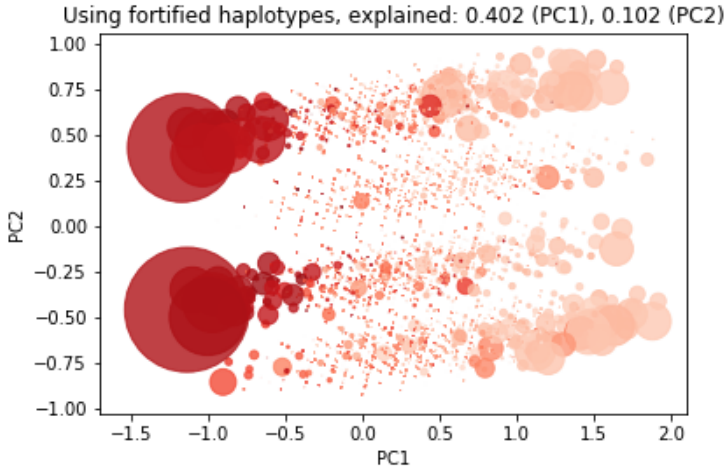


Figure 2.4: The first two principal components of the fortified haplotype space and the proportions of variance they explain. The first PC separates the haplotypes by their colour score, the rest of the visible structure is largely due to the set of all haplotypes on twelve loci forming a twelve-dimensional hypercube. Interestingly, the large, dark red haplotypes seem to be much more similar to each other than the smaller, more variable pale haplotypes.

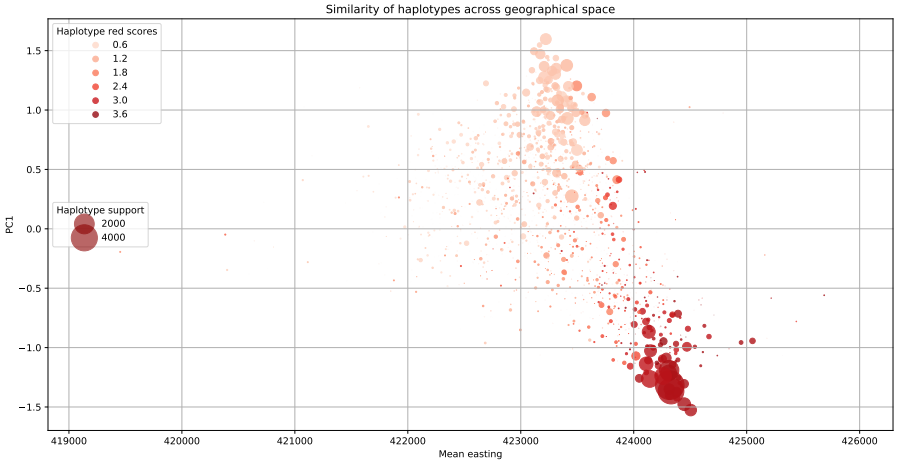


Figure 2.5: PC1 vs. Easting. PC1 separates the dark phenotypes from the rest, but does not capture the structure of the pale phenotypes. The many haplotypes of small support with intermediate phenotypes which are typical for the the hybrid zone core are projected to the middle of the PC1 as well.

different PCs, as this would only cause "seeing" the hypercube from a different angle. One could train the PCA on a subset of the most common haplotypes, which would lead to several of the smaller haplotypes projecting on top of each other, as the directions that they differ in would not be represented in the PCs at all.

As seen in figure 2.5, PC1 also separates the haplotypes along the east-west axis (figure 2.5), although it does better job in separating the dark red haplotypes from the pale haplotypes than in differentiating between the pale haplotypes.

2.2.1 Haplotype genealogy

Another space to project the haplotypes into to look at the variation and relationships between them is to use a "genealogy" (figure 2.6): considering the four main branches (A, B, C and D), out of the two largest ones, D seems to be exclusively dark red and A mostly pale and patchy. (The phenotype and the mean phenotype of "pale" recessive haplotypes will largely depend on the haplotypes they are paired with and therefore on the haplotypes growing near them.) The two smaller branches B and C seem to be a mixture with low support and hence, with low certainty about the average colours. When we relate the haplotypes with their mean geographical positions (figure 2.7), we see that although the two smallest branches only constitute 4 and 11 haplotypes respectively, they contain 7 out of 27 most "central" haplotypes and thus the "central" haplotypes are overrepresented in the two small branches. This, together with overall appearance of the haplotypes (mostly zeroes on one side and mostly ones on the other) in the two small branches suggests that they could mainly consist of recombinants.

Overall, the haplotypes of similar mean red scores do cluster together as expected, even if not due to linkage of the SNPs to the causal loci, then due to geographical distribution of the two different phenotypes. Another plus is, that we can indeed get rid of the hypercubical structure by displaying the haplotypes in a dendrogram. However, although recombination quite certainly plays an important role in the system and seems to be indicated in the two small branches (B and C), simple approaches such as neighbour joining or hierarchical clustering of the haplotypes do not inherently capture it. Therefore, more complex methods would be needed to understand the biological relationships between the haplotypes.

2.2.2 Possible effects of other variables on estimating haplotype frequencies via EM

The version of the EM used for estimating haplotype frequencies does not take into account geographical proximity and linkage (genetic proximity). In reality, haplotypes carried by two nearby plants are more likely to combine in a genotype than those that grow on the opposite ends of the hybrid zone. Similarly, closest SNPs might be a better predictor of some focal SNP than the ones at the other side of the haplotype (discussed in section 3.2.2). Not including these two factors is not necessarily a shortcoming, but it might be worthwhile to investigate and understand the effects of this decision on the results.

The effect of geographical proximity

One could argue that haplotypes carried by two nearby plants are more likely to combine in a genotype than those that grow on the opposite ends of the hybrid zone. For purpose of investigating the effect of geographical distance in the model, we are going to compare the results from training on the subset of genotypes of plants only located in the *core* and then, on all genotypes from the hybrid zone in Planoles (all genotypes in *numerous* dataset). The *core* individuals here are defined as all plants from a central and most "colourful" 2-kilometre-long section of the 10-kilometre-long hybrid zone. For simplicity, the value of Easting was chosen to be between $422,000$ and $424,000$ metres for the *core* individuals (to see why that this is also a good choice of a "colourful" region see figure fig:freq-geo-rosel-types). This *core* dataset comprises of about 58% of all plants in *numerous*.

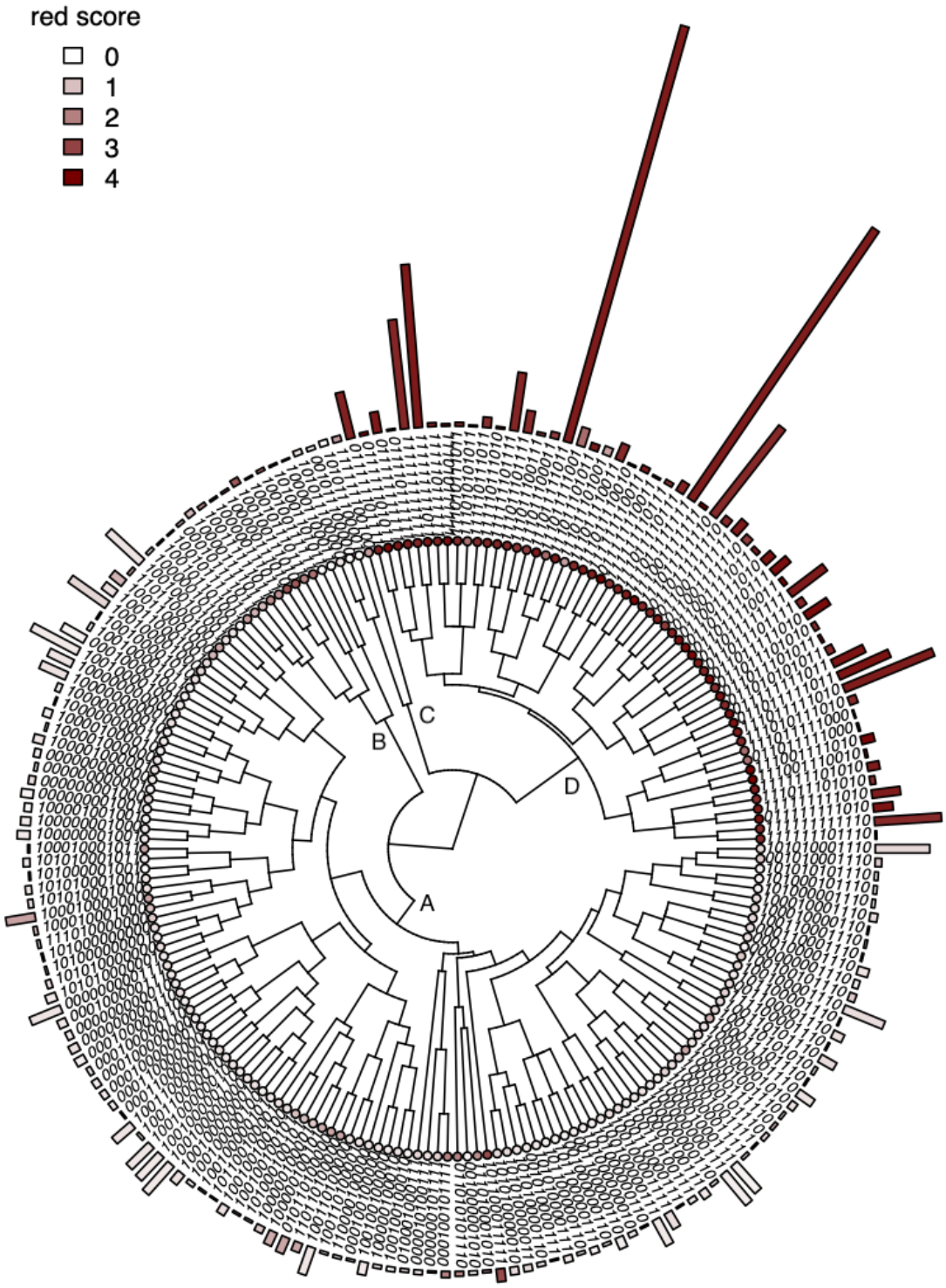


Figure 2.6: The "genealogy" of haplotypes calculated by agglomerative hierarchical clustering algorithm (unweighted pair group method with arithmetic mean), with their estimated mean red score and estimated support denoted by the length of the bars, ranging from 10 to over $5,060$ copies. Haplotypes estimated to be present in less than 10 copies are omitted for simplicity. The four main branches are denoted with letters A, B, C and D. While the two largest branches, A and D are mostly homogeneous in their red score, branches B and C are more heterogeneous. Please note that this is not a true genealogy due to disregarded recombination, most apparent in the branches B and C.

Figure 2.7: The genealogy of haplotypes related to the mean positions of the haplotypes and their mean red score on the background of all plants in the population denoted with black dots. The genealogy itself is the same as the one in figure 2.6, the red "core" line is at the same position as the core in figure 2.8 and the pink line shows the most "central" haplotypes. The four main branches are denoted with letters A, B, C and D. While the two largest branches, A and D are pointing almost exclusively to mean positions in the west and east of the hybrid zone, respectively, branches B and C are more heterogeneous and "central" haplotypes are over-represented in them.

To obtain the *core*-based estimates, we first need to calculate the *core*-based contribution matrix is $\mathbf{A^t}_{core}$. It can be estimated using the same EM algorithm described in section 2.1.2, just using the numbers of genotypes in the *core* $\mathbf{n_g}_{core}$ instead of using the original numbers of genotypes $\mathbf{n_g}$ coming from the *numerous* dataset. The *core*-based frequency estimates can then be calculated as

$$\mathbf{p}_{core} = \frac{\mathbf{n_h}_{core}}{2n_{core}} = \frac{\mathbf{A^t}_{core}\mathbf{n_g}_{core}}{2n_{core}}$$

When we compare the results of the EM on *core* individuals to those on all individuals on *numerous*, we can observe a predictable pattern (first subfigure in figure 2.8): The frequency of rare haplotypes ("small") with very central mean positions well inside the *core* are now estimated to be present in much higher frequencies (yellow). This is understandable, as individuals carrying these rare and very central haplotypes are mostly growing in the core, but not outside the core, so their proportion in the *core* haplotypes simply *is* larger. On the contrary, the very dark red haplotypes present mostly in the plants on the east side of the hybrid zone were creating a large proportion of the *numerous* haplotypes, but many of the plants carrying them are now excluded from the analysis by only considering the individuals from the core. Hence, their estimated frequencies must have decreased (dark blue).

All in all, this first comparison is a good exercise for checking whether the algorithm is giving reasonable results. However, it does not shed much light on the effect of ignoring the geographical distance, as most of the changes seem to be explainable by differing numbers of genotypes between the two datasets, $\mathbf{n_g}$ and $\mathbf{n_g}_{core}$. Therefore, we decided to compare the original estimates from *numerous* (calculated using *numerous*-based contribution matrix $\mathbf{A^t}$ and numbers of genotypes in *numerous* $\mathbf{n_g}$) with frequencies calculated using *core*-based contribution matrix $\mathbf{A^t}_{core}$ and numbers of genotypes in *numerous* instead of the numbers of genotypes in *core* (second subfigure in figure 2.8):

$$\mathbf{p}_{core-matrix} = \frac{\mathbf{n_h}_{core-matrix}}{2n} = \frac{\mathbf{A^t}_{core}\mathbf{n_g}}{2n},$$

where $\mathbf{n_g}$ might be missing some elements, as some genotypes are not found in the core and are not accounted for in the *core*-based matrix. In such case, the n is adjusted accordingly to be the sum of the remaining elements of the $\mathbf{n_g}$.

Quite encouragingly, the results from this comparison (second subfigure in figure 2.8) show close to no differences to the frequencies of haplotypes with mean positions **in the *core***, as all points seem uniformly violet. To support what we see with numbers: the difference from the original frequencies in haplotypes with mean positions in the *core* range from -10^{-3} to 10^{-2}. This can be a lot compared to the originally predicted frequencies, especially for haplotypes that were originally predicted at very low frequencies (for example 10^{-66}, i.e. most likely not present at all). In haplotypes that were originally predicted at at least 10^{-4} the maximum change is less than two-fold. More precisely, for the haplotypes with mean positions in *core* that were originally predicted at at least 10^{-4}, the change relative to originally predicted frequencies (ratio) ranges from -1 to 1.92.

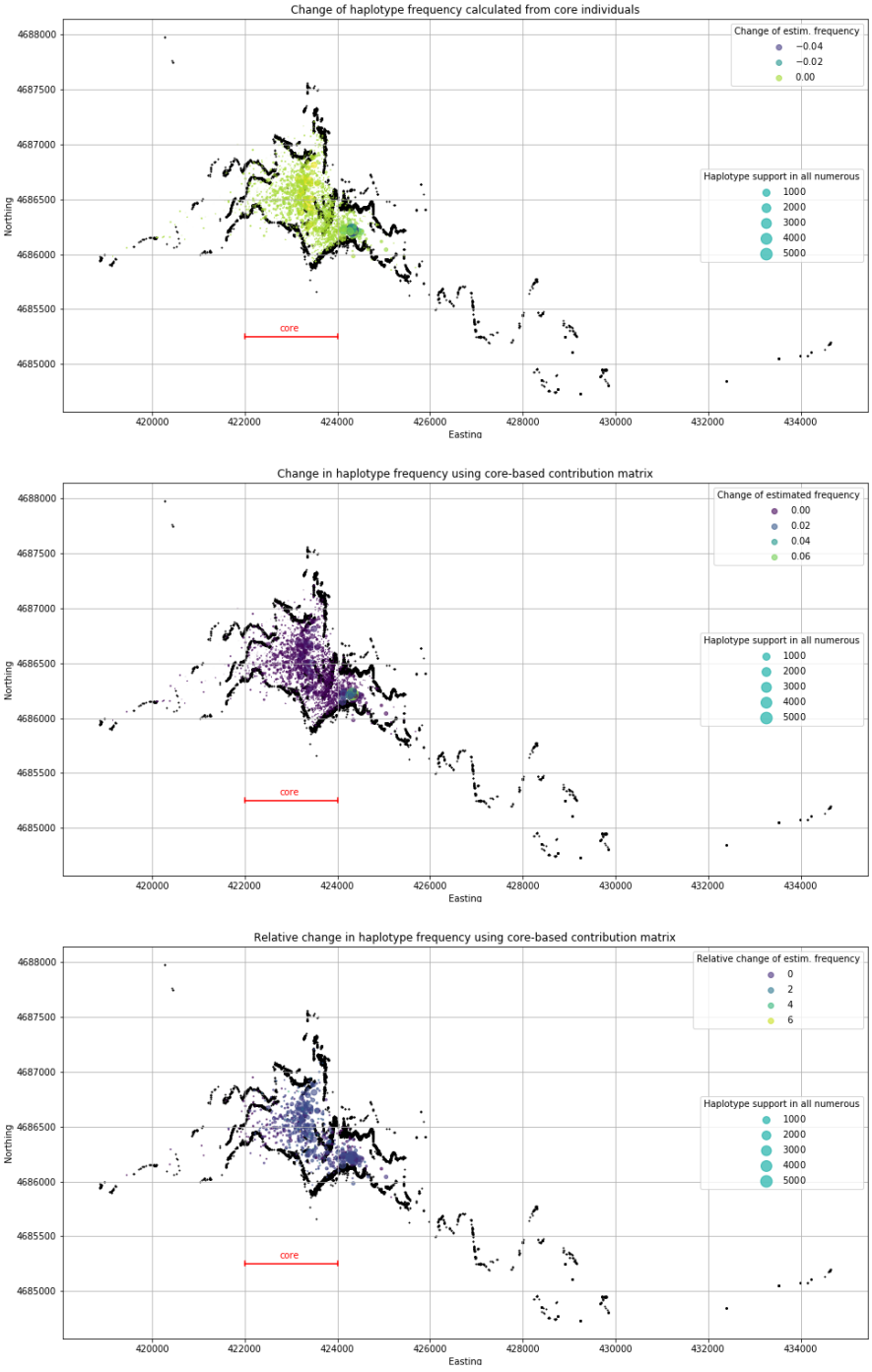
The largest absolute differences in frequency estimates seems to be in some of the most common haplotypes associated with high red scores and mean positions **on the east from the *core***. More precisely, the frequencies rose by values as high as 2 to 7%, which may seem a lot. However, considering only the haplotypes that were originally predicted at at least 10^{-4}, the relative changes to originally predicted frequencies range from -1 to 1.41. The effect of linkage is discussed in section 3.2.2.

2.3 Conclusions

Firstly, and quite congruent with our expectations, haplotypes with weak support are projected into the middle of both haplotype and geographical space, clustering together in the two smallest branches out of the four main branches of the haplotype genealogy dendrogram (figures 2.6 and 2.7). This makes sense, because the core of the hybrid zone is exactly the place where dark and pale haplotypes meet and recombine.

A more suprising finding is that there seems to be asymmetric diversity in the haplotypes, demonstrated across all of the figures in this chapter (especially 2.5 and 2.6). In particular, there seem to be many more distinct pale haplotypes with intermediate estimated support, positioned in the western part of the hybrid zone. This contrasts with fewer dark haplotypes with very large estimated support, present in the eastern part of the hybrid zone. This would suggest that the *Ros/El* region playing crucial role in anthocyanin production and distribution is conserved in the eastern, "dark" side of the hybrid zone, whereas in the west it is not as important and hence more prone to variation. In other words, *there are only a few ways to be red, but many ways not to be*. We study this phenomenon in connection to states of genes *Rosea* and *Eluta* in chapter 3.

Figure 2.8: Three different ways of comparing EM-estimated haplotype frequencies from the *core* to the original estimates from the *numerous* dataset. The colourful discs represent haplotypes with colours denoting the difference between the new estimates and the original ones. For easy comparison, the positions and sizes of the markers correspond to the haplotype mean positions and support in the *numerous* dataset and they are the same as in figure 2.3. The red line shows the range of Easting defining the *core* dataset.



CHAPTER 3

Haplotypes in genetic context

In this chapter, which is an extension of chapter 2, we study structure of the haplotypes in relation to the *Rosea* and *Eluta* loci and their association with the flower phenotype, as these two closely-linked loci have been shown to explain the majority of variation in the amount of floral anthocyanin.

In particular, we study haplotype structure in the context of the four main *Rosea*/ *Eluta* (*Ros*/*El*) types and correlations between the individual markers based on their positions relative to *Rosea* and *Eluta*. Out of the four main *Ros*/*El* types, the first two are parental types: *ROSel*, associated with magenta-coloured *Antirrhinum majus pseudomajus* growing on the Eastern side of the hybrid zone and *rosEL*, associated with yellow-coloured *Antirrhinum majus striatum* growing in the West. The second two are recombinant types *ROSEL* and quite rare *rosel*. We will focus on their geographical location, their diversity and the phenotype they carry. To do that, we consider 12-SNP genotypes of $\sim 22,300$ plants and 24-SNP genotypes of $\sim 6,500$ plants in the *Ros/El* region from Planoles hybrid zone population. We call these datasets *numerous* and *long* and they are described in section 2.1.1. In addition to raw genotypes we use a *genotype-to-haplotype* distribution obtained from Expectation Maximisation algorithm and the haplotype frequencies in the sample as described in 2.1.2. More details about all SNPs in the region of interest can be found in table 2.1 in chapter 2.

The three main problems we study in this chapter are the following:

The *Eluta* phenotype problem The effects of *Eluta* described in 1.4.2 have mostly been studied in combination with recessive *rosel* haplotype. Here we use the large, real-world dataset to study the effects of Eluta in more complex situations.

Differences in haplotypic diversity As we already observed in chapter 2, there seems to be much less diversity in haplotypes associated with dark red phenotypes than there is in those associated with pale phenotypes. Rather than in relation to red scores, here we study haplotypic diversity in relation to the four Ros/El types.

Distribution of the four Ros/El types The *rosel* and *ROSEL* recombinants should be generated from the *ROSel*/*rosEL* heterozygotes at the same rate. However, there seems to be deficit of the *rosel* haplotypes. Here we use the large dataset to study this phenomenon.

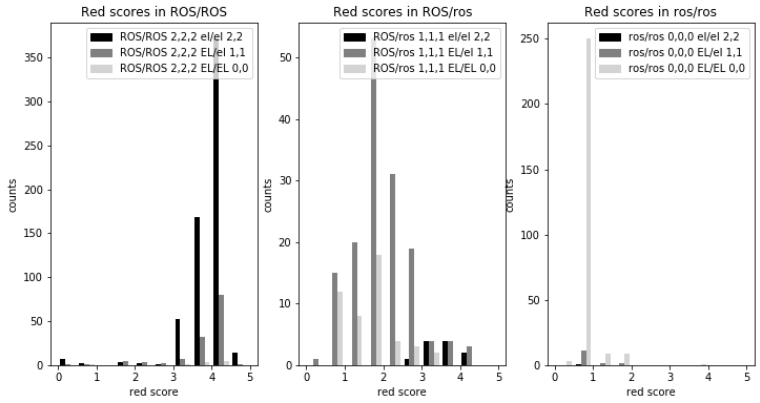


Figure 3.1: Histograms of red scores for plants with various genotypes at *Rosea* and *Eluta*. The only noticeable effect a dominant version of *Eluta* has seems to be in heterozygous *ROS/ros* plants.

3.1 Flower colour phenotypes

In this section we study the relationship between genotype, haplotype and the amount and distribution of magenta coloured anthocyanin pigment in flowers and compare it to the previous findings (summarised in section 1.4.2). The manual "red scores" used here to describe the amount and distribution of magenta pigmentation according to figure 4.5 in [67]. They range from 0 to 5 by (usually) 0.5. We often refer to full magenta flowers as "dark red" or "full red". These would have scores of 3.5 or higher. Flowers with intermediate red score values of 2 - 2.5 would be referred to as "patchy" and those with low red score values of 0 - 1 would be called "pale".

These manual scores are clearly ordered (although there may be slight disparities between scores by different scorers), but they are not necessarily referring to equidistant states (of pigment amounts, or anything else). Therefore, statistics assuming continuous distribution are not suitable and we use non-parametric tests whenever possible.

3.1.1 Red score distribution in genotypes

Firstly, let us look at the relationship between the diploid genotypes and phenotypes. We extracted genotypes "RosEl_All" and manual red scores from the "Final_Geno_Ecol" file and summarised the results in a histogram (figure 3.1).

As expected, in homozygous *ros/ros* plants the presence of a dominant *Eluta* allele does not seem to have a measurable effect on the distribution of anthocyanin, as flowers of these plants are very low in athocyanin in general.

For heterozygous *ROS/ros* plants, the phenotypes behave mostly as expected, i.e. the presence of at least one dominant *Eluta* allele shifts the flower colour from full red phenotypes ($3-5$, black bars) to the typical non-homogeneous *Eluta* range (1.5 - 2.5) and even below (0.5 - 1). (One-sided Mann-Whitney U test with continuity correction p-value: 1.57×10^{-7}.) Although there does not seem to be a visible difference between the distribution of homozygous *EL/EL* and heterozygous *EL/el* individuals, the red score distribution of homozygous *EL/EL* individuals is actually shifted to the left. (One-sided Mann-Whitney U test with continuity correction p-value: 1.24×10^{-3}.)

Perhaps most surprising is therefore the panel with *ROS/ROS* homozygotes, where the presence of at least one dominant *Eluta* allele **does not** cause the clear shift towards the non-homogeneous *Eluta* range (compared to the *ROSel/ROSel* phenotypes) as in the *ROS/ros* case. Although one could argue that there seems to be a little shift and the values $1.5 - 2.5$ are more common in *EL/el* individuals than they are in *el/el* individuals although there are less *EL/el* individuals overall, the difference in red score distribution is not statistically significant in *EL/el* heterozygotes (one-sided Mann-Whitney U test with continuity correction p-value: 0.242) nor in *EL/EL* homozygotes (one-sided Mann-Whitney U test with continuity correction p-value: 0.162), both compared to *el/el* homozygotes.

3.1.2 Red score distribution in haplotypes

Unfortunately, not all plants in the *numerous* dataset are genotyped at all of 24 loci in table 2.1. However, we can still look at one marker in *Rosea 1*, two markers at *Rosea 3* and one marker at *Eluta* at about $22,300$ plants. To begin simply, we choose the marker 3 in *Rosea 1* to indicate the *Rosea* state and 17, the only available marker in *Eluta* to indicate the *Eluta* state. We translate the state at these two SNPs to dominant and recessive *Rosea* and *Eluta* alleles as indicated in table 3.1 and we will use this grouping into the four *Ros/El* types throughout this chapter.

state at *Rosea 1*	state at *Eluta*	*Ros/El* haplotype
0	0	rosEL
0	1	rosel
1	0	ROSEL
1	1	ROSel

Table 3.1: States at Rosea 1 and Eluta translating to the four main Ros/El haplotype types.

Firstly, we calculated marginal distributions of the red scores for each haplotype: for a given haplotype, each of the red scores $(0.5, 1, 1.5, ...4.5)$ has a weight corresponding to how often and how strongly this haplotype was indicated at the given red score (see figure 3.2).

Unsurprisingly, the haplotypes of parental *Ros/El* types seem to have mostly unimodal red scores distribution corresponding to the parental types. On the other hand, recombinant haplotypes seem to have red distributions with several peaks: rosel at 0.5 and 3.5 and ROSEL at 0.5, at 2 - 2.5 and at 4, making it roughly trimodal. And, indeed: the resulting phenotype heavily depends on the other haplotype, which is much more variable in the hybrid zone core, where most of the recombinants are located.

The next logical step is therefore to look at the red score distributions of the **pairs of haplotypes** indicated to form a genotype together (see figure 3.3). The presence of distinctly coloured horizontal stripes in the third row of panels indicates that, in combination with various ROSel haplotypes, several of the recombinant *ROSEL* haplotypes seem to consistently yield full red phenotypes. This is consistent with the findings in figure 3.1) rather than with the literature-predicted intermediate phenotypes. There also seems to be one aberrant *ROSel* haplotype yielding different phenotypes than all the other *ROSel* phenotypes.

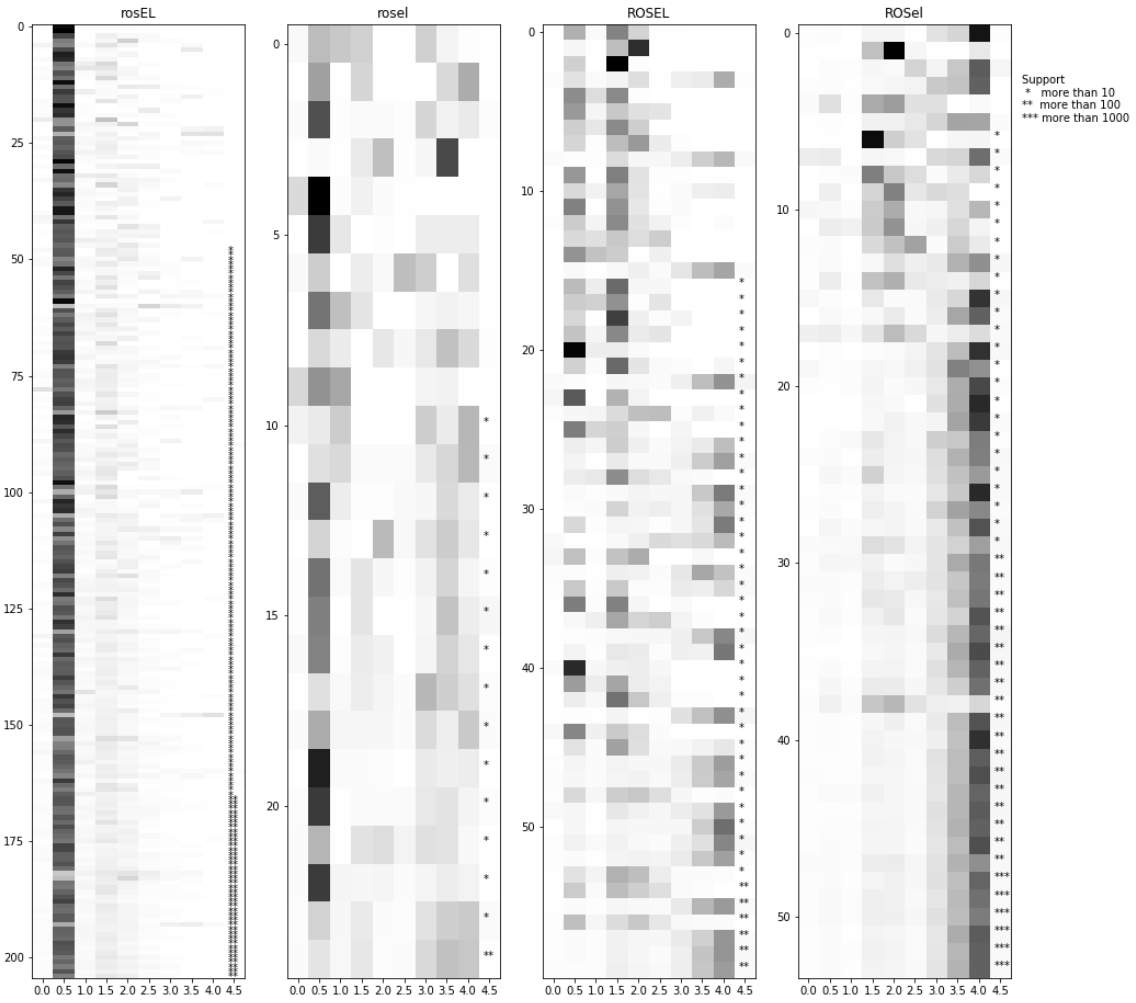


Figure 3.2: Marginal distributions of the red scores for the four *Ros/EI* haplotype types. Each row inside one of the four panels corresponds to a different haplotype, with red scores indicated on the horizontal axis. The darkness of a single field shows the proportion of the given haplotype (row) being indicated in a plant with the given red score (column), with darker shades corresponding to higher proportions. Each of the four panels contains data on haplotypes belonging to one of the four *Ros/EI* types (rosEL, rosel, ROSEL and ROSel) indicated on the top of the panel. Inside of the panel, the haplotypes are ordered according to predicted frequencies inside the population, with the most common types on the bottom. Their predicted number of copies is also designated with a number of stars (see the legend in top right). Only the haplotypes predicted at 5 copies or more are plotted here. We can see that although plants carrying *ROSEL* haplotype (third column) would be dominant both in *Ros* and in *EI* no matter what other haplotype the *ROSEL* would be combined with, we do not see the single expected peak of the distribution in the intermediate values between 1.5 and 2.5 for all the *ROSEL* haplotypes. Many of them have are predominantly full red (4) and some of them are almost exclusively pale (0.5), which is in disagreement with the null hypothesis 1.4.2 that the amount of anthocyanin can be predicted simply by two loci with two alleles each.

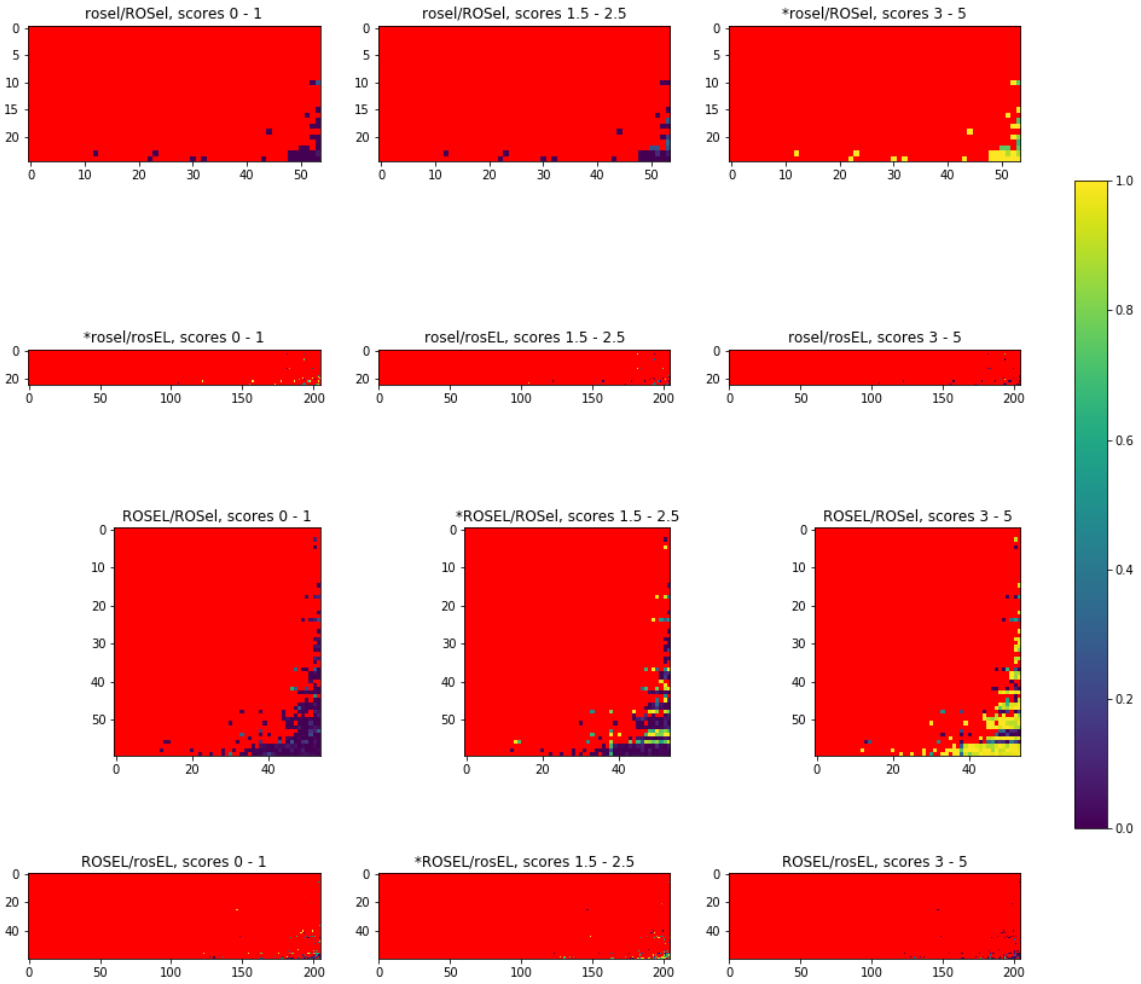


Figure 3.3: Red score distribution for pairs of haplotypes falling to different *Ros/El* categories. Each of the three columns of the colourful panels corresponds to one red score group: pale $(0-1)$, intermediate $(1.5-2.5)$ and dark $(3-5)$. Each of the four rows of the panels corresponds to a different cross of recombinant and parental haplotype: *rosel/ROSel*, *rosel/rosEL*, *ROSEL/ROSel* and *ROSEL/rosEL*. Inside a panel, each position corresponds to a given couple of haplotypes, with recombinant haplotypes in rows and parental haplotypes in columns. This position square is coloured according to the proportion of the plants (whose genotypes were formed by this given couple of haplotypes) belong to the one of three red score categories. Therefore, the values of the given position across the three "red score category" columns sum up to one. Only the haplotypes predicted at five copies or more are plotted here. The haplotypes are in the same order as in the figure 3.2. Haplotype pairs predicted to form genotype of less than two plants are shown in red, same as missing data.

3.2 Patterns in haplotypic diversity

In chapter 2 we observed that haplotypes associated with pale phenotype are much more diverse than haplotypes associated with dark red phenotype. Here, we test whether this observation translates to the *Rosea* and *Eluta* alleles carried by these haplotypes, i.e. whether rosEL haplotypes coming from the pale parental population of *A. m. striatum* are more diverse than ROSel haplotypes coming from the dark red (full magenta) parental population of *A. m. pseudomajus*.

Ros/El type	total copies	haps at > 1	haps at > 5	haps at > 10
rosEL	$15,600$	314	205	157
rosel	786	61	25	15
ROSEL	$3,591$	126	60	44
ROSel	$20,556$	71	54	48

Table 3.2: Haplotypes of Ros/El types, their total predicted numbers in population, and numbers of unique haplotypes predicted as present above given thresholds: at more than one, more than five and more than copies in the sample.

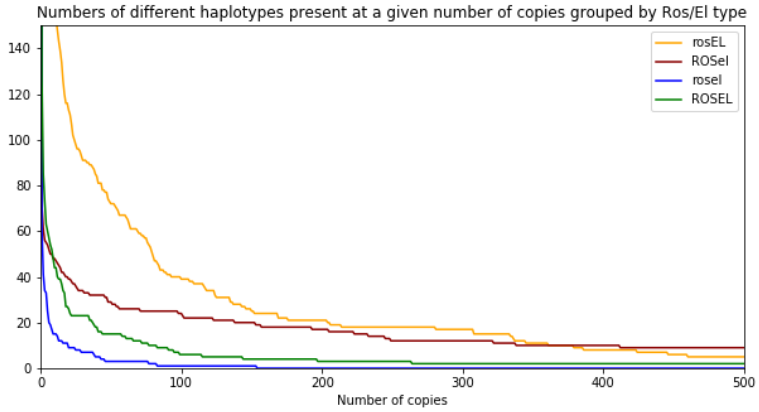


Figure 3.4: Variability of haplotypes of different *Ros/El* types in the hybrid zone. Number of different haplotypes found in population at least given number of copies.

To find out more, we calculated the predicted numbers of haplotypes from $\sim 22,300$ plants in the *numerous* dataset belonging to the four main Ros/El types (table 3.2). Our predictions have been confirmed: although $ROSel$ haplotypes are predicted to be in the population at considerably more copies than $rosEL$, there seem to be many more different unique $rosEL$ haplotypes (table 3.2). This is depicted in more detail in figure 3.4 showing decreasing diversity within the four Ros/El haplotype groups with increasing threshold. Here we can see that not only does the line for rosEL haplotypes cross with the line for ROSel haplotypes (orange and red, respectively) showing an imbalance in haplotypic diversity vs. the number of their predicted copies between rosEL and ROSel, but the same applies even to ROSEL and ROSel haplotypes. This imbalance in diversity between ROSel and rosEL (and between ROSel and ROSEL) haplotypes has several possible explanations, our two favourite ones follow.

First comes the selection hypothesis. One could argue, that since *Rosea* and *Eluta* play crucial roles in anthocyanin pigment production and distribution, these loci will be more constrained by selection in fully magenta-coloured population of *A. m. pseudomajus* carrying the *ROSel*

allele. (The importance of being fully magenta in a region where all flowers are full magenta could be conferred by pollinators, for example.) In other words: there are only a couple of ways to be red, while there are many more ways not to be.

Alternatively, this phenomenon could be explained by demographic history. In that case, the rosEL subspecies (*A. m. striatum*) would come from an old and diverse population of *A. majus*, while the ROSel subspecies (*A. m. pseudomajus*) would descend by a recent expansion from a small and more uniform population, perhaps as a mutant carrying the full magenta coloration.

We could distinguish between these two hypotheses by analysing patterns of haplotypic diversity in other genetic regions. A straightforward idea would be to focus on a region responsible for differences in yellow pigmentation between the two subspecies. However, the main locus involved in the process is characterised by a deletion in yellow *A. m. striatum* and therefore we cannot use it for this purpose.

The higher diversity in ROSEL haplotypes compared to ROSel haplotypes can be still driven by the diversity in the *Eluta* side of the ROSEL haplotypes, coming from the diverse rosEL haplotypes. This hypothesis can be tested by identifying location of diversity hotspots in the set of ROSEL haplotypes.

The numbers of copies of rosel haplotypes are so low, that it is difficult to tell whether there is any conflict.

3.2.1 The *rosel* deficit

To study the geographical distribution of the four Ros/El types, we plotted estimated numbers of copies from each of the four Ros/El types grouped into bins across the Easting, both in absolute numbers and as proportions (figure 3.5). Interestingly, there seems to be deficit of the *rosel* haplotypes, although recombinant haplotypes *rosel* and *ROSEL* should be generated from the *ROSel/rosEL* heterozygotes at the same rate. The effect seems to be especially strong on the yellow side of the hybrid zone, where there seems to be much less *rosel* than *ROSEL* haplotypes in most of the bins. Moreover, while *ROSEL* seems to be present throughout the hybrid zone, *rosel* is basically nonexistent to the east from the core.

At the moment, we can only speculate about the reason for the *rosel* haplotype deficit we observe in the data. For example, there may be a selection against plants carrying the *rosel* haplotype, or there could have been a historical event leading to this inequality, which is now just being perpetuated.

To understand whether the *rosel* deficit could arise just by chance under symmetric conditions, Nick Barton ran a preliminary simulation based on code from rotation of Andrea Mrnjavac. In this stepping stone model simulation, two subspecies come into contact in the middle of the hybrid zone. The hybrid zone is modelled as 20 demes with $N = 1000$ haploid genomes each, with migration $m = 0.5$ between the two neighbouring demes and it was run for 100 generations. The haploid genomes on 24 SNPs uniformly spread over 1 cM, with two SNPs "*Rosea*" and "*Eluta*" at 0.25 and 0.75 cM selected with strength $s = 0.08$ in opposing directions in the two halves of the hybrid zone. The results have shown that the recombinants are concentrated in the middle of the hybrid zone, symmetrically decreasing in numbers (and proportions) as we move through demes from the centre outwards from the very beginning and this pattern is only stabilised across the generations. The numbers of the two types of recombinants ("*ROSEL*" and "*rosel*") are roughly equal throughout the generations.

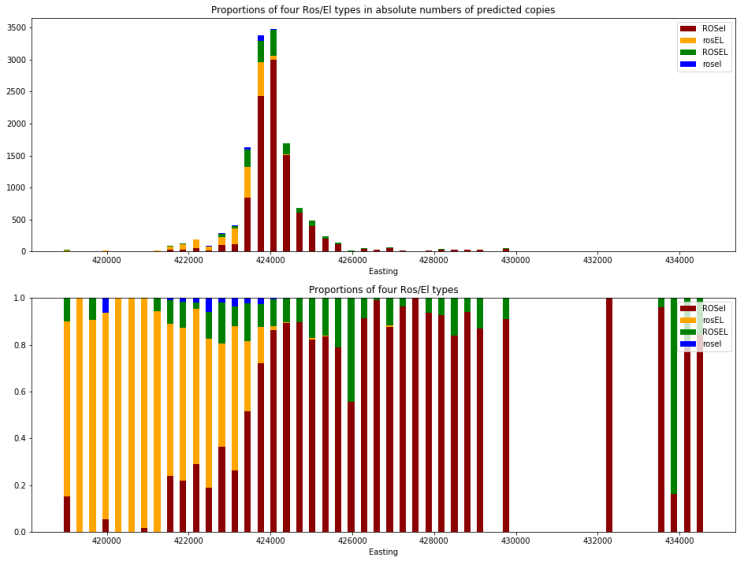


Figure 3.5: Proportions of four Ros/El types across the Easting. In real numbers of predicted copies in the population in the upper panel and normalised to sum up to one in the lower panel.

To determine how unusual seeing as extremely low proportion of *rosel* recombinants as we actually see in our data under symmetric conditions (i.e. a *p-value*) is beyond the scope of this thesis, but we suggest it as an interesting future project. One could run this simulation 500 - 1000 times and calculate the proportion of as extreme or more extreme results. If this proportion lower than e.g. 0.05, the simple symmetric hypothesis could be rejected. However, the positions of SNPs in these simulations would ideally be adjusted to match the real SNP positions.

3.2.2 Linkage in *numerous* haplotypes

In this section we study linkage between the 12 SNPs in the *numerous* dataset. Their positions relative to the first SNP in *numerous* as well as their positions are depicted in figure 3.6, more information on the SNPs can be found in table 2.1.

Most phasing methods estimate haplotypes locally, as it is believed that in general, closest SNPs are better predictors of a focal SNP state than more distant ones. Phasing locally also makes sense in context of very long genotypes and small sample sizes. However, it also means that genetic correlations caused by other factors than distance may be distorted and the results can be distance-biased.

Since our sample comprises more than $22,000$ genotypes on only 12 SNPs in our region of interest, we can afford to use Expectation Maximisation algorithm considering all possible 2^{12} haplotypes at once. This way, our estimates are not biased by only considering the SNPs nearby. Furthermore, this means that we can investigate linkage disequilibrium from the estimated haplotype frequencies without worrying about using the same positional information twice.

We calculated the correlation coefficient, r, between two loci (table 3.3), which is coefficient

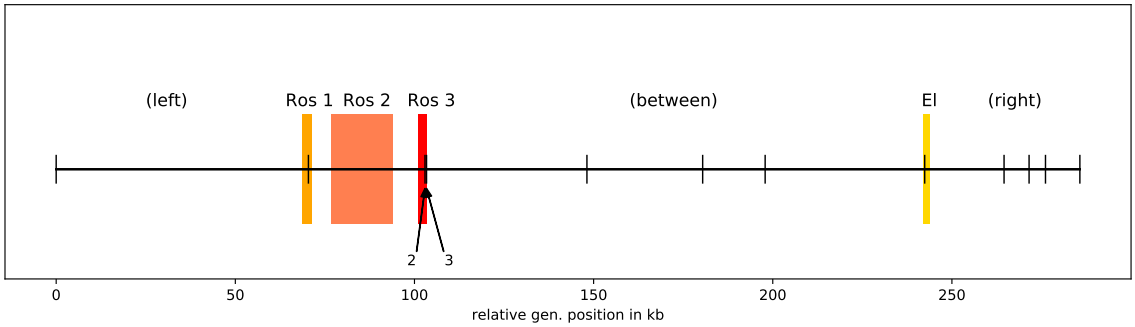


Figure 3.6: Positions of 12 *numerous* SNPs relative to *Rosea* and *Eluta* genes. Positions of *Rosea* and *Eluta* genes according to [40] are denoted by colourful rectangles. Genomic positions on chromosome 6 in kilobases are given relative to position of the first SNP in the *long* dataset. SNPs outside of *Rosea* and *Eluta* are described as "left", "between" (*Rosea* and *Eluta*) and "right" (in parentheses). SNPs 2 and 3 are very close to each other in *Ros 3*.

id	0	1*	2*	3*	4	5	6	7*	8	9	10	11
0	0.00	-0.01	0.01	-0.05	0.12	0.05	0.04	0.13	0.06	-0.07	-0.03	0.01
1*	-0.01	0.00	0.75	-0.24	0.65	0.83	0.69	0.85	0.41	0.04	0.36	0.28
2*	0.01	0.75	0.00	-0.37	0.70	0.78	0.60	0.82	0.43	0.12	0.43	0.24
3*	-0.05	-0.24	-0.37	0.00	-0.13	-0.21	-0.09	-0.22	-0.03	-0.01	-0.11	-0.09
4	0.12	0.65	0.70	-0.13	0.00	0.69	0.62	0.70	0.36	-0.01	0.28	0.18
5	0.05	0.83	0.78	-0.21	0.69	0.00	0.65	0.83	0.38	0.07	0.38	0.33
6	0.04	0.69	0.60	-0.09	0.62	0.65	0.00	0.85	0.40	0.01	0.31	0.30
7*	0.13	0.85	0.82	-0.22	0.70	0.83	0.85	0.00	0.39	-0.02	0.34	0.25
8	0.06	0.41	0.43	-0.03	0.36	0.38	0.40	0.39	0.00	0.05	0.31	0.22
9	-0.07	0.04	0.12	-0.01	-0.01	0.07	0.01	-0.02	0.05	0.00	0.25	0.02
10	-0.03	0.36	0.43	-0.11	0.28	0.38	0.31	0.34	0.31	0.25	0.00	0.19
11	0.01	0.28	0.24	-0.09	0.18	0.33	0.30	0.25	0.22	0.02	0.19	0.00

Table 3.3: Correlation coefficients between all pairs of SNPs in the *numerous* dataset. The SNP 1 is located in *Ros1*, 2 and 3 are in *Ros3* and SNP 7 is in *El*. For more information on SNPs please refer to table 2.1.

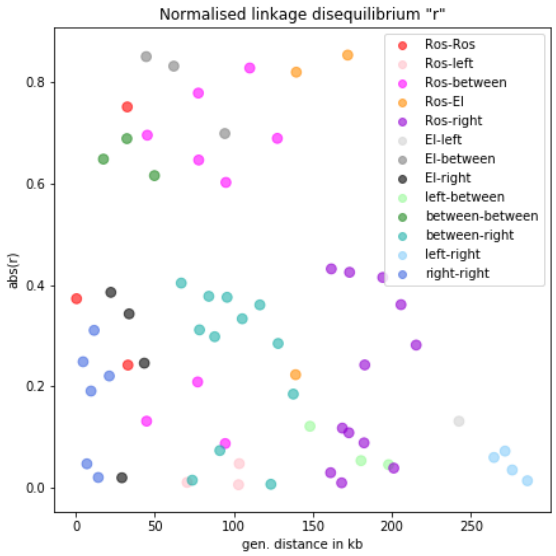
of linkage disequilibrium normalised by all four allele frequencies

$$r_{AB} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} = \frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}.$$

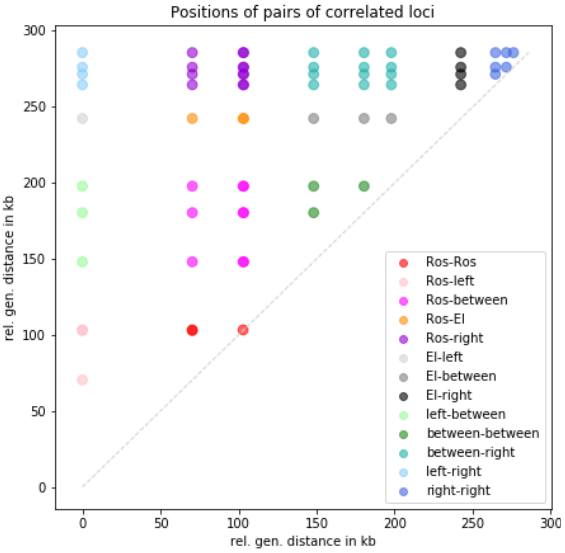
In our case, allele A means that there is a 1 at the position A of the haplotype, AB means that there is 1 at both loci A and B. We can calculate r using the estimated haplotype frequencies $\mathbf{p}$, such that $p_A = \sum_{i \in Q_A} p_i$ and $P_{AB} = \sum_{i \in Q_{AB}} p_i$, where Q_A is a set of all haplotypes that have a 1 at position A and Q_{AB} is a set of all haplotypes that have a 1 at both loci A and B. Since the choice of denoting alleles with 0 and 1 is arbitrary, the sign of r does not have any information value and we only plot the absolute value of r (figure 3.7a).

If all linkage disequilibrium was explicable by genetic distance, we should see decreasing values of $|r|$ with increasing genetic distance. However, this is not completely apparent from the data. The values of $|r|$ vs. genetic distance seem to form two, horizontally separated clouds.

Figure 3.7: Normalised linkage disequilibrium between all pairs of SNPs in the *numerous* dataset.



(a) Absolute values of correlation coefficient r between all pairs of SNPs in the *numerous* dataset vs. their distance in kilobases. Pairs containing SNPs in *Ros* are in shades of red (pink, red, orange, magenta and purple), pairs not containing *Ros* SNPs which have the *El* SNP are in shades of grey and all other correlations are in shades of blue and green. All outlier pairs containing a *Ros* SNP contain the SNP 3 in *Ros3*.



(b) Positions of pairs of correlated SNPs from the previous plot. Only one disc per pair is plotted, all above the dashed line denoting theoretical pairs with equal positions. The positions are in kilobases, relative to the left-most SNP of the *numerous* dataset. The SNPs number 2 and 3 (both in Ros 3) lie only 400 base pairs apart, so all pairs of shape SNPX&SNP2 and SNPX&SNP3 appear to be on top of each other and hence the corresponding discs appear to be a more saturated colour.

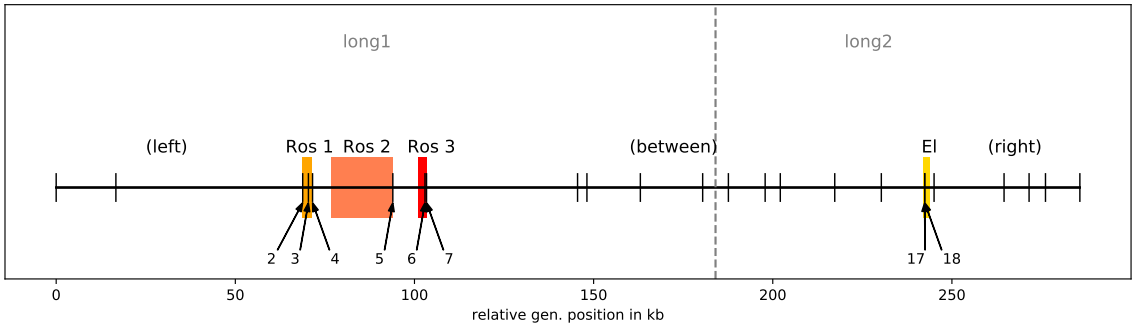


Figure 3.8: The position of SNPs relative to *Rosea* and *Eluta* genes in the *long* dataset. Positions of *Rosea* and *Eluta* genes according to [40] are denoted by colourful rectangles. Genomic position on chromosome 6 in kilobases is given relative to position of the first SNP in the *long* dataset. The dashed grey line separates the 24 SNPs of the *long* dataset into 12 SNPs of *long1* (around *Rosea*) dataset and 12 SNPs of *long2* (around *Eluta*) dataset. SNPs outside of *Rosea* and *Eluta* are described as "left", "between" (*Rosea* and *Eluta*) and "right" (in parentheses). SNPs 6 and 7 are very close to each other in *Ros 3* and SNPs 17 and 18 are just 14 base pairs apart in *Eluta*. SNP 4 is very close to, but just outside of *Ros 1*.

In the upper cloud with $|r| > 0.5$ there are only correlations between *Ros*, *El* and *between* SNPs (SNPs $2, 3, ...7$). On the other hand, all correlations involving SNPs from *left* or *right* group of SNPs seem to have an absolute value below 0.4.

Interestingly, there are also two of the three *Ros-Ros* correlations, one *Ros-El* and a couple of *Ros-between* correlations. When we consult the table 3.3 with values of r, we can see that all values of $|r|$ from the lower cloud in fact come from correlations with SNP number 3, i.e. a SNP from third exon of *Ros 3*. A similar analysis for more detailed, twice as dense *long* genotypes can be found in section 3.3.2.

3.3 A more detailed view

In this section we use the more detailed genotypes on 24 SNPs from the *long* dataset available for about $6,500$ plants (described in 2.1.1) to study the genetic neighbourhood of the *Rosea* and *Eluta* genes (SNP positions relative to *Rosea* and *Eluta* are in figure 3.8). Here, we split the *long* into two halves and we use the information on the two halves of the genotypes separately: *long1* consisting of first twelve SNPs including six SNPs from the three *Rosea* genes (therefore we call this half *"around Ros"*, available for $6,790$ plants) and *long2*, second twelve SNPs including two SNPs from *Eluta* (and hence, we call this half *"around El"*, available for $6,989$ plants). Both *long1* and *long2* include six SNPs present in the *numerous* dataset. Each plant genotyped for *long* loci is also genotyped for all of the *numerous* loci, as these form a proper subset. This makes it possible to relate these more detailed genotypes to the sparser, yet more widely measured *numerous* genotypes. More detailed information on the SNP positions, their positions relative to genes and their presence in *numerous* dataset can be found in table 2.1.

Figure 3.9 shows numbers of plants with complete *long1* and *long2* genotypes available compared to the available *numerous* genotypes. The first message we get is that the numbers of *long1* and *long2* genotypes are very similar. This is because it is largely the same plants, but due to failed measurements of SNPs in first half, or the other, different numbers of plants

were excluded from the two datasets. We can see that there are never more *long* genotypes available than there are *numerous* genotypes. This is because *numerous* SNPs are subset of the *long* SNPs, therefore each plant for which the *long* genotype is available there is also a *numerous* genotype. And most importantly, around Easting $424,000$, just where most plants are genotyped, there is a big dip in the proportion of plants with *long* genotype. Since this area is enriched in the dominant *ROS* allele (see figure 3.5), this might lead to different haplotype frequency estimates. However, when we compare the real proportions of the genotypes grouped by their Ros/El types (table 3.4), the proportions seem to be very similar.

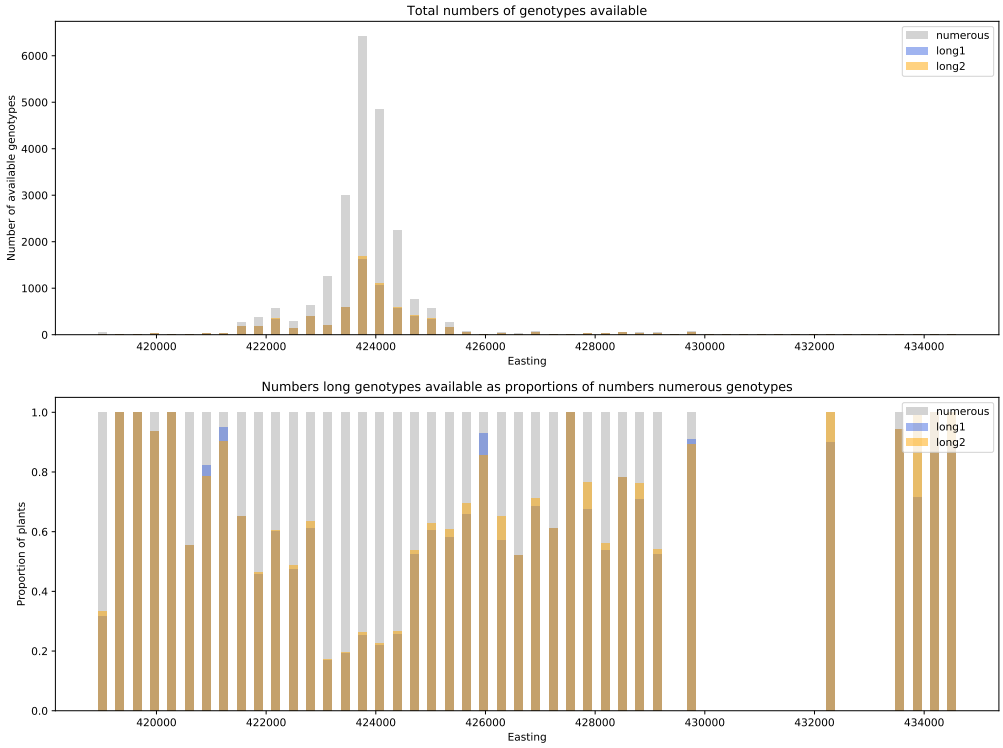


Figure 3.9: Easting positions of plants with complete *long1* and *long2* genotypes available compared to the plants with *numerous* genotypes available, above in total numbers, below as proportions of numbers of *numerous* genotypes available. Around Easting $424,000$, there is a significant dip in the proportion of plants with *long* genotypes. This region is enriched in the dominant *ROS* allele (see figure 3.5).

3.3.1 Long haplotypes: frequencies, colour and geography

The haplotype sets and their frequencies were obtained, as before, using the Expectation Maximisation Algorithm (section 2.1.2) with prior information (as described in section 7) exploiting the information we already had from calculating the same for *numerous* dataset. Just to double-check, we compared the frequencies of common *partial* haplotypes present in both *numerous* and *long1* and *long2* datasets in figure 3.10. Here, the common partial haplotypes are haplotypes on *long* SNPs 0, 3, 6, 7, 9 and 11 (SNPs 0 - 5 in *numerous*) for *long1* and haplotypes on SNPs 13, 17, 20, 21, 22 and 23 (SNPs 6 - 11 in *numerous*) for *long2*. To obtain the frequencies of partial haplotypes, all haplotype frequencies with the given values at the fixed positions are summed up.

Rosea	Eluta	Counts in numerous	Counts in long	Proportions in numerous	Proportions in long
0	0	5700	1752	0.26	0.27
0	1	406	94	0.02	0.01
0	2	24	4	0.00	0.00
1	0	1122	313	0.05	0.05
1	1	4002	1029	0.18	0.16
1	2	346	63	0.02	0.01
2	0	199	57	0.01	0.01
2	1	2325	722	0.11	0.11
2	2	7877	2438	0.36	0.38

Table 3.4: Counts and proportions of *Ros/El* genotypes in *numerous* and in *long*. The genotypes are grouped by values at SNPs in *Rosea 1* (number 1 in *numerous*, number 3 in *long*) and *Eluta* (number 7 in *numerous*, 17 in *El*). The proportions of the genotypes in the two datasets seem to be very similar, if anything, there are slightly fewer plants homozygous for parental genotypes in *numerous*.

For example, *if v denotes a position where a certain haplotype set does have a value and $*$ a position where it does not, a common partial haplotype for first haplotype set $*vv$ and second haplotype set $vv*$ would be $*v*$. To get the frequency of e.g. partial haplotype $*1*$ in the first set we add frequencies of haplotypes $*10$ and $*11$ estimated in the $*vv$ haplotype set, to get the frequency of the same partial haplotype in the second set we add frequencies of haplotypes $01*$ and $11*$estimated in the $vv*$ haplotype set.*

Interestingly, the differences between *long*-derived and *numerous*-derived common partial haplotype frequencies around *Eluta* are much lower than those around *Rosea*, which can be explained by the sampling bias captured in figure 3.9: In the *long* dataset, the dominant *ROS* allele is undersampled, since proportion of plants with *long* genotypes is the lowest around Easting $424,000$, which has the most sampled plants and $70 - 99\%$ (depending on Easting) of them are *ROS*.

When we plot the *long1* and *long2* haplotypes on the map to analyse the their mean geographical position, mean red score and support in the sample, we get a figure similar to that for the *numerous* haplotype (figure 2.3). However, while the *long1* haplotypes around the *Rosea* locus seem to be even more orderly, clearly separated by the Easting into pale haplotypes in the West and dark ones in the East, with even fewer haplotypes of intermediate redscores, *long2* haplotypes do not follow this pattern. This suggests that the haplotype at and around *Rosea* is much more important for the red score than that around the *Eluta* locus and *Eluta* haplotypes might not be as tightly tied to a geographical position (if it was, its mean red score would be much more determined by the prevalent *Rosea* type, which is, in turn, very strongly determined by the geographical position and hence the map would have to be much more homogeneous than it is now). Also, while both *Rosea* and *Eluta* haplotypes seem to have darker and larger haplotypes concentrated on the East side, the *Eluta* haplotypes are much more diverse, with lower support, which is apparent particularly when comparing the left part of the two plots.

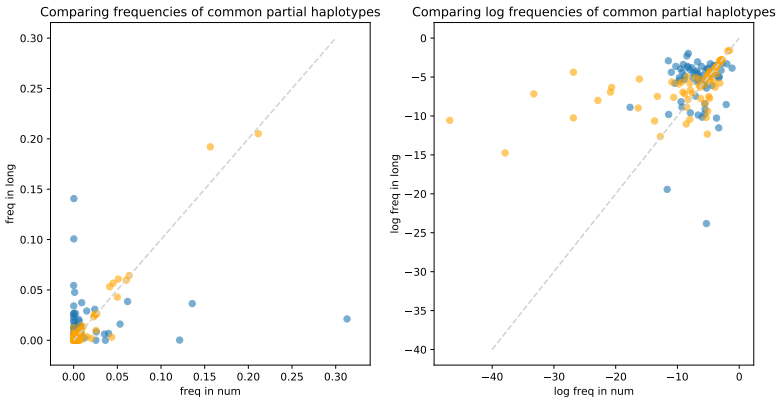


Figure 3.10: Comparison of frequencies of common partial haplotypes in *long1* and *long2*, both in normal and in a natural logarithmic scale. Each disc represents one partial haplotype, the partial haplotypes from the first half of the SNPs (relevant to *long1*) are shown in blue, the partial haplotypes from the second half of the SNPs (relevant to *long2*) are shown in orange. The coordinates of the discs represent their frequency in *numerous* and in the relevant *long* (either *long1* or *long2*) haplotype sets. The grey dotted line denotes positions of all partial haplotypes present at equal frequencies in *long* and *numerous* haplotype sets.

3.3.2 Haplotype structure and linkage

Finally, we analyse the linkage within the *long1* and *long2* haplotypes and the haplotype structure in relation to the four Ros/El types. Similar to figure 3.7 in section 3.2.2, where we calculated correlation between pairs of markers in *numerous* haplotypes, this time we calculated the correlation between loci within *long1* and within *long2* in figures 3.13 and 3.14 and in tables 3.5, 3.6, respectively. (Since the haplotype structures and their frequencies were estimated via two *separate* runs of the Expectation Maximisation Algorithm for the two disjoint sets *long1* and *long2*, calculating correlation between pairs of loci across these two datasets, i.e. for pairs of SNPs where one comes from *long1* and the other from Ilong2 is not possible in the same way.)

We classify the 24 SNPs in *long* into five non-overlapping categories, from SNPs with lowest ID number to the highest, as shown in figure 3.8: *left* (markers upstream from *Rosea* genes), *Ros* (SNPs within *Rosea* genes, further divided into *Ros1*, *Ros2* and *Ros3*), *between* (area between *Rosea* and *Eluta*), *El* (the two SNPs within *Eluta*) and *right* (all markers to the "right" from *Eluta*). There are two pairs of SNPs very close to each other: two in *Ros3* (430 base pairs apart) and two in *El* (14 base pairs apart). The gaps between three *Ros1* SNPs are approximately $1,600$ and $1,200$ base pairs.

As before in *numerous* (figure 3.7), correlations within protein-coding genes in *Ros* and *El* and in pairs with *between* SNPs tend to be high, with an exception of pairs involving second marker in *Ros3* (here SNP number 7, in *numerous* 3). Out of these, the highest correlations are between pairs of SNPs within *Ros1* and within *El* (as expected) and between *Ros1* and *Ros2*. This may be because the *Ros1* and *Ros3* regions are well conserved, but also a by-product of SNP selection: The SNPs in *Ros1* and *Ros3* may have been chosen such that they are predictive of the subspecies (always with one allele common in *A. m. pseudomajus*, the other in *A. m. striatum*) and strongly clinal.

Consistent with findings in *numerous*, correlations between pairs of SNPs containing *left* and

id	0	1	2*	3*	4*	5*	6*	7*	8	9	10	11
0	0.00	-0.08	0.01	0.00	0.04	0.02	0.00	-0.04	-0.01	0.11	0.01	0.06
1	-0.08	0.00	0.08	0.05	0.04	0.05	0.09	0.09	-0.04	0.06	0.00	0.04
2*	0.01	0.08	0.00	0.98	0.96	0.97	0.72	-0.20	-0.35	0.68	0.58	0.81
3*	0.00	0.05	0.98	0.00	0.98	0.99	0.74	-0.26	-0.36	0.65	0.60	0.83
4*	0.04	0.04	0.96	0.98	0.00	0.96	0.73	-0.28	-0.37	0.63	0.59	0.81
5*	0.02	0.05	0.97	0.99	0.96	0.00	0.76	-0.28	-0.37	0.63	0.62	0.84
6*	0.00	0.09	0.72	0.74	0.73	0.76	0.00	-0.40	-0.52	0.70	0.52	0.78
7*	-0.04	0.09	-0.20	-0.26	-0.28	-0.28	-0.40	0.00	0.26	-0.13	-0.21	-0.21
8	-0.01	-0.04	-0.35	-0.36	-0.37	-0.37	-0.52	0.26	0.00	-0.36	-0.20	-0.34
9	0.11	0.06	0.68	0.65	0.63	0.63	0.70	-0.13	-0.36	0.00	0.43	0.70
10	0.01	0.00	0.58	0.60	0.59	0.62	0.52	-0.21	-0.20	0.43	0.00	0.64
11	0.06	0.04	0.81	0.83	0.81	0.84	0.78	-0.21	-0.34	0.70	0.64	0.00

Table 3.5: Correlation coefficients between all pairs of SNPs in the *long1* dataset. Markers 2, 3 and 4 are located in *Ros1*, 5 is in *Ros2* and 6 and 7 are in *Ros3*. For more information on SNPs please refer to table 2.1.

right SNPs tend to be also quite low. However, this trend seems to be more pronounced in the first half of the SNPs (figure 3.13a) than in the other (figure 3.14a), which could be, to some extent, explained by smaller genetic distances in *long2* dataset, especially in the subset of four *right* markers. A notable difference with the results from *numerous* seems to be in dispersal of correlations within *between* SNPs: While the three *between* SNPs in *numerous* were quite highly correlated (all three pairs have $r > 0.6$), in *long1* the absolute values of correlations among four *between* SNPs range from 0.20 to 0.70 and in *long2* the correlations among five *between* SNPs range from 0.03 to 0.87. Where do the differences come from? We should bear in mind that the sample of plants in the *long* dataset is different than the one in *numerous*. Firstly, it could contain a different proportion of recombinant genotypes. However, we know that the differences should be minimal, since overall, the proportions of the diploid Ros/El genotypes are very similar (exact numbers table 3.4). Secondly, the *long* dataset contains less genotypes, so the estimated correlations may simply be less precise. To verify this, we compared correlations in *numerous* and *long* for all pairs available for both datasets (the pairs that have one SNP in *long1* and the other in *long2* cannot be used) in figure 3.12. However, the available values are really similar regardless of value of r, so it seems that values of correlation can really be trusted.

Haplotype structure in the Ros/El context

All whole haplotypes can be divided into four basic Ros/El categories based on their states at the *Rosea* and *Eluta* genes. Out of the four Ros/El types, two are parental: ROSel typical for magenta coloured *Antirrhinum majus pseudomajus* and rosEL typical for yellow *Antirrhinum majus striatum* and two recombinant types *ROSEL* and *rosel*. For simplicity, we decided to define these four groups based on states in two SNPs: the state in Rosea is represented by SNP in *Ros 1* (here SNP number 3, number 1 in *numerous*) and *El* (here SNP number 17, number 7 in *numerous*), as described in table 3.1. These SNPs were chosen because they are genotyped in most of the plants in *numerous* and *long* datasets and they correlate with the parental types. However, states of other SNPs in *Rosea* and *Eluta* are available and although the correlation of pairs of SNPs within the *Rosea* and *Eluta* genes is rather high, it is not

id	12	13	14	15	16	17*	18*	19	20	21	22	23
12	0.00	0.12	0.22	-0.26	0.03	0.17	0.17	0.21	0.04	0.01	0.13	0.07
13	0.12	0.00	0.87	-0.43	0.55	0.85	0.85	0.88	0.38	0.06	0.34	0.31
14	0.22	0.87	0.00	-0.36	0.32	0.58	0.58	0.57	0.26	0.09	0.22	0.26
15	-0.26	-0.43	-0.36	0.00	-0.22	-0.41	-0.41	-0.55	-0.27	-0.09	-0.26	-0.16
16	0.03	0.55	0.32	-0.22	0.00	0.87	0.87	0.63	0.34	-0.18	0.18	0.22
17*	0.17	0.85	0.58	-0.41	0.87	0.00	0.99	0.76	0.37	0.01	0.36	0.25
18*	0.17	0.85	0.58	-0.41	0.87	0.99	0.00	0.76	0.36	0.01	0.36	0.25
19	0.21	0.88	0.57	-0.55	0.63	0.76	0.76	0.00	0.40	0.05	0.39	0.33
20	0.04	0.38	0.26	-0.27	0.34	0.37	0.36	0.40	0.00	0.10	0.34	0.23
21	0.01	0.06	0.09	-0.09	-0.18	0.01	0.01	0.05	0.10	0.00	0.25	0.03
22	0.13	0.34	0.22	-0.26	0.18	0.36	0.36	0.39	0.34	0.25	0.00	0.20
23	0.07	0.31	0.26	-0.16	0.22	0.25	0.25	0.33	0.23	0.03	0.20	0.00

Table 3.6: Correlation coefficients between all pairs of SNPs in the *long2* dataset. Markers 17 and 18 are located in *El*. For more information on SNPs please refer to table 2.1.

absolute.

To better understand the structure of haplotypes in the four Ros/El types, we estimated haplotype frequencies (around *Ros* and around *El* separately) within plants homozygous at these four Ros/El types (with two fixed SNPs each) using their *long* genotypes, since whole 24-SNP-long haplotypes are not available. Therefore, we should keep in mind, that we are only using genotypes of $\sim 4,400$ plants homozygous at the two loci and we are omitting about $2,000$ plants, thus losing a lot of information on recombinant haplotypes, which are quite rare and even rarer as homozygotes. To get the haplotype frequencies, we multiplied the relevant genotype counts by contribution matrices coming from the EM algorithm (more details in section 2.1.2). Then we plotted all haplotypes above a certain threshold in order of their decreasing frequency (figure 3.15).

Quite interestingly, ROSel haplotypes are still about half as variable as the rosEL haplotypes both around *Ros* and around *El*, requiring about twice as many haplotypes to cover both the first 50% and the first 90% of all plants in the category. This is in accordance with our previous findings where ROSel haplotypes were much less diverse than rosEL haplotypes.

We can hypothesise that this may be due to selection on the *Rosea* genes, as these tend to be less diverse than SNPs around *Eluta*, albeit this alone would be a slightly unfair comparison, considering that there are six SNPs inside the *Rosea* genes and only two *Eluta* SNPs.

However, it seems that while in *ROSel* haplotypes, SNPs number 13, 16, 18, 22 and to a large extent 19 and 20 seem to be conserved around fixed SNP number 17, in *rosEL* haplotypes mostly just the SNP number 18 (and to some extent also SNPs number 15 and 19) are conserved. It is quite interesting that SNP number 15 seems to be conserved in the dominant EL allele in rosEL context, but not in the ROSEL context.

3.4 Summary

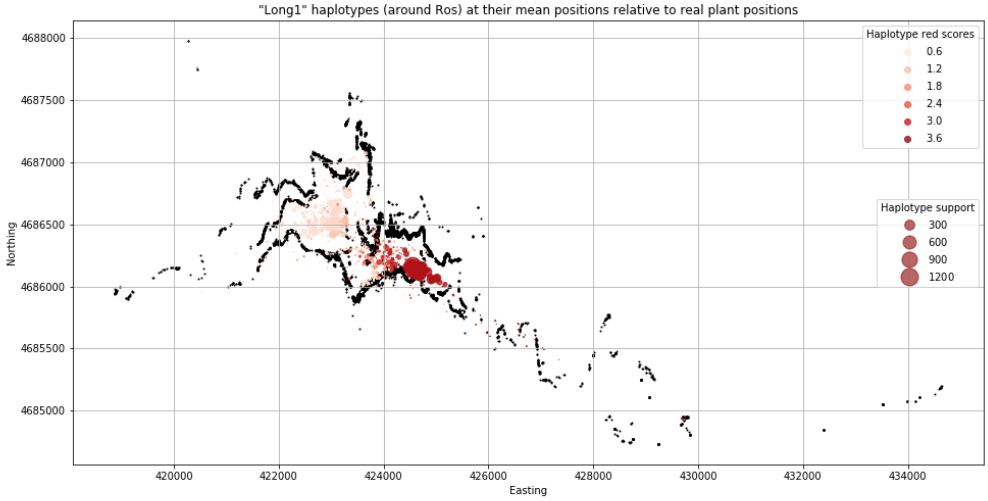
This chapter yields several confirmed hypotheses, but also a couple of unexpected conclusions.

Firstly, we found differences from previously described genetic effects on flower colour. As expected and described in 1.4.2, Eluta is not demonstrated in pale (ros/ros) plants. However, from what we see it seems that it is only manifested in ROS/ros plants and the effect in dark red ROS/ROS plants is almost negligible (figure 3.1). However, there exists a certain group of ROSEL haplotypes which tends to consistently yield intermediate red score values in combination with various ROSel haplotypes (figure 3.3).

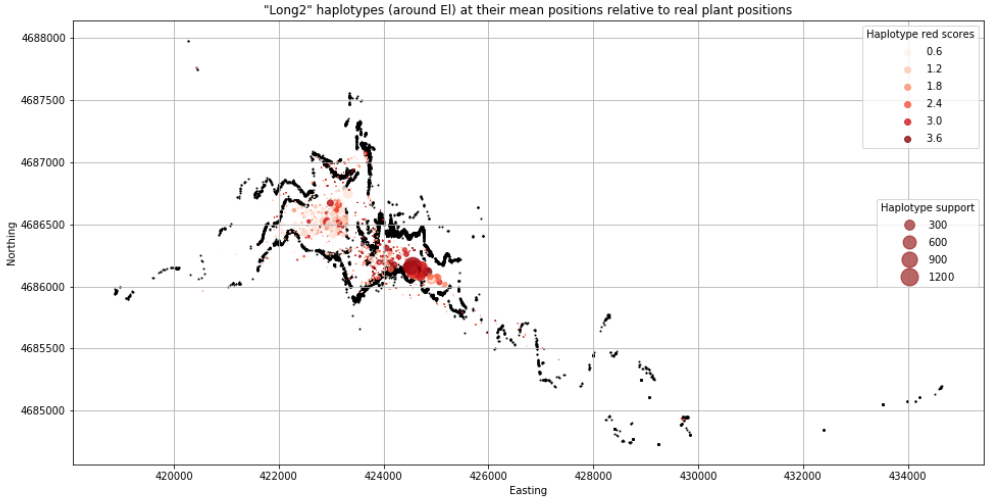
As we saw in chapter 2, despite overall higher numbers of dark red plants collected, the "dark" haplotypes seem to be less diverse than the "pale" haplotypes. Here we confirmed, that this phenomenon translates into the *Rosea* and *Eluta* context and there are fewer unique *ROSel* haplotypes, although there are more *ROSel* haplotypes indicated in the sample (table 3.2 and figure 3.4). We offered two explanations, one based on selection and the other one on demographic history, suggesting ways to identify the more likely one. We attempted to answer this question comparing haplotype structure around *Rosea* vs. that around *Eluta* using more detailed haplotypes in the *Ros*/*El* region, but this is not trivial to do, largely due to the choice of genotyped SNPs in the sample.

Finally, the *rosel* and *ROSEL* recombinants should be generated from the *ROSel*/*rosEL* heterozygotes at the same rate. However, there seems to be deficit of the *rosel* haplotypes, especially on the yellow side of the hybrid zone. We suggested a method to determine the significance of this finding via simulations.

Figure 3.11: *Long1* and *long2* haplotypes at their estimated mean geographical positions (discs in shades of red) plotted on the background of all plants (in the *numerous* dataset) at their real positions (in black). The shades of red denote the estimated mean red score associated with each haplotype and its support in the data is represented by the size of the marker. Figure 2.3 is a similar plot for *numerous haplotypes*.



(a) *Long1* haplotypes (around Ros) at their mean positions relative to real plant positions. This plot is very similar to that for *numerous* haplotypes (figure 2.3), except it is even more "orderly", with pale and dark haplotypes geographically cleanly separated by the Easting, with very few haplotypes with intermediate mean red score values, which are all rare (small) and concentrated in the middle.



(b) *Long2* haplotypes (around El) at their mean positions relative to real plant positions. The haplotypes around *Eluta* are not geographically divided into groups of haplotypes with homogeneous mean red scores, which suggests that the state of these SNPs is not associated with red score as the SNPs around the *Rosea* locus. Moreover, haplotypes around *Eluta* are more diverse (although common haplotypes around *Eluta* still seem to be concentrated on the right side) and seem to be also much less connected to a certain geographical position than haplotypes around *Rosea*.

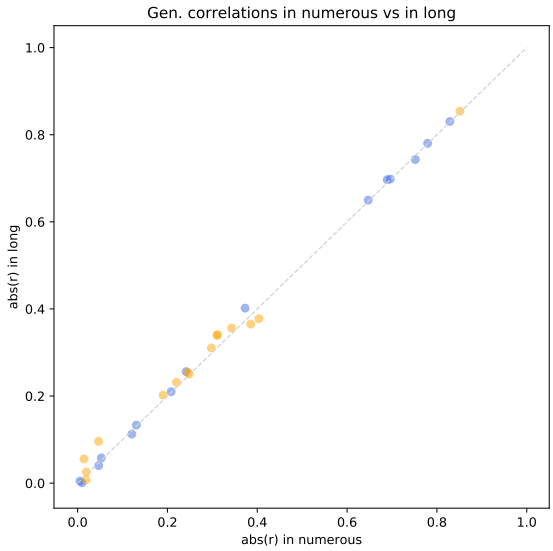
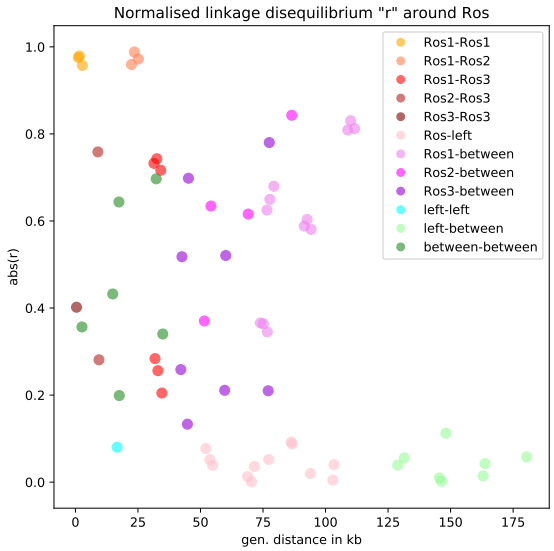
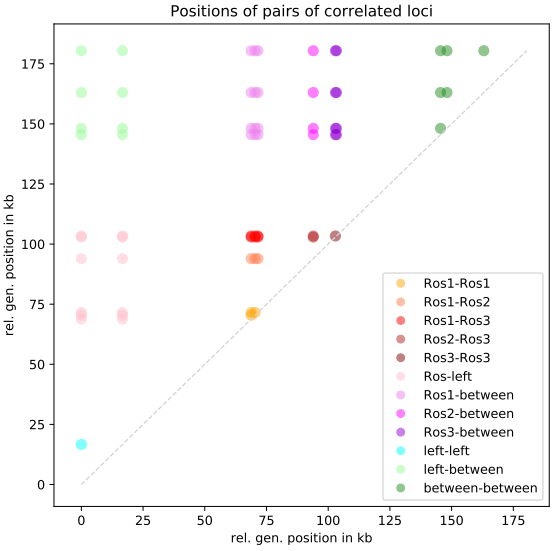


Figure 3.12: Comparison of normalised linkage disequilibrium calculated in *numerous* with that calculated in *long*. Values from *long1* are shown in blue, those from *long2* in orange. We can see that the differences are minimal regardless of value of r.

Figure 3.13: Normalised linkage disequilibrium between all pairs of SNPs in the *long1* dataset

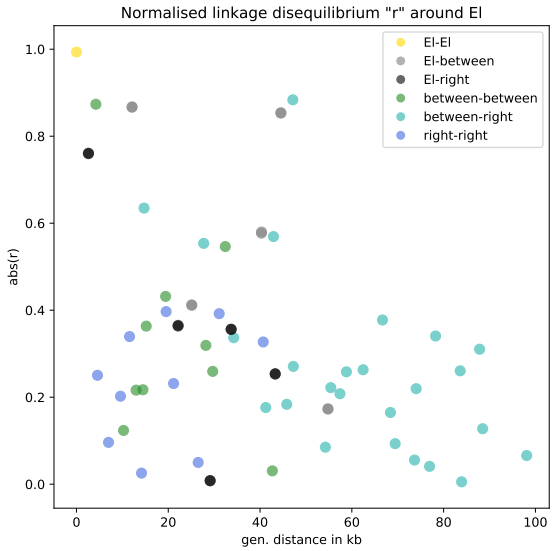


(a) Absolute values of correlation coefficient r between all pairs of SNPs in the *long1* dataset vs. their distance in kilobases. Pairs containing SNPs in *Ros* are in shades of red (pink, red, orange, [illegible]Ps are in shades of blue and green.

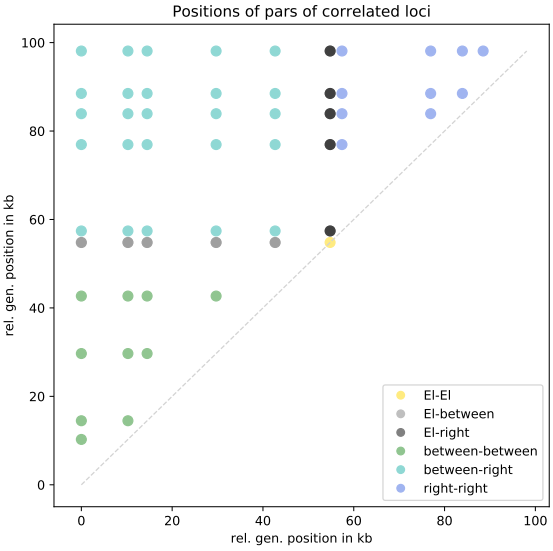


(b) Positions of pairs of correlated SNPs from the previous plot. Only one disc per pair is plotted, all above the dashed line denoting theoretical pairs with equal positions. The positions are in kilobases, relative to the left-most SNP of the *long1* dataset. The SNPs number 2,3 and 4 (all in Ros 1) lie within $1,200$ base pairs, so all pairs including these SNPs are somewhat close. The SNPs number 6 and 7 (all in Ros 3) lie 400 base pairs from each other, hence all pairs involving SNPs 6 and 7 appear to be on top of each other and therefore the corresponding discs appear to be a more saturated colour.

Figure 3.14: Normalised linkage disequilibrium between all pairs of SNPs in the *long2* dataset



(a) Absolute values of correlation coefficient r between all pairs of SNPs in the *long2* dataset vs. their distance in kilobases. The disc denoting correlation between the two *Eluta* SNPs is yellow, other airs containing SNPs in *Eluta* are grey and black and pairs not containing *Eluta* SNPs are in shades of blue and green. The two *Eluta* loci only about 30 base pairs apart seem to be almost



(b) Positions of pairs of correlated SNPs from the previous plot. Only one disc per pair is plotted, all above the dashed line denoting theoretical pairs with equal positions. The positions are in kilobases, relative to the left-most SNP of the *long2* dataset. The SNPs number 17 and 18 (here 5 and 6) lie within 30 base pairs from each other, so all pairs appear to be on top of each other and therefore the corresponding discs appear to be a more saturated colour.

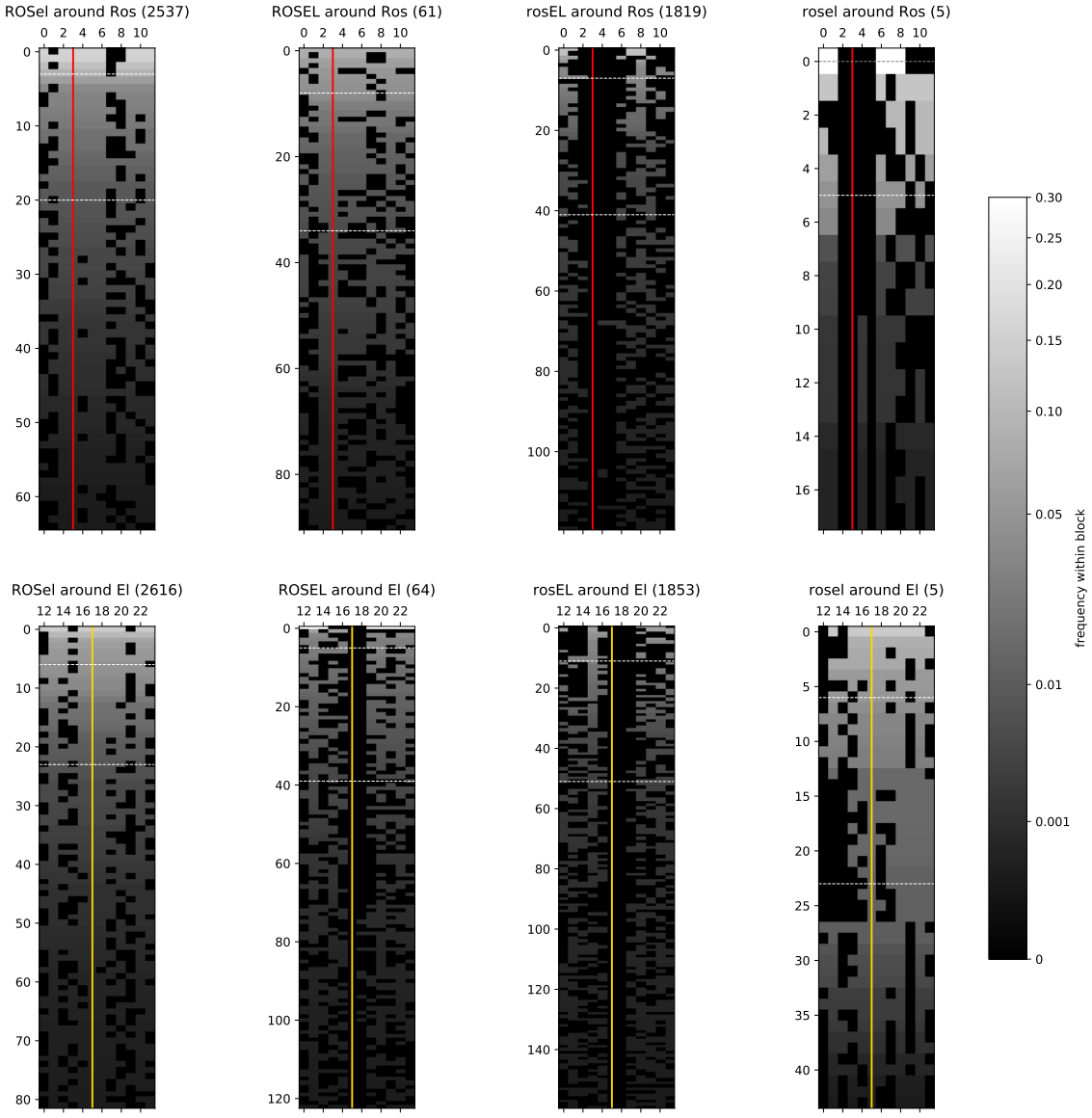


Figure 3.15: Haplotype structure in plants homozygous for the four Ros/El types. The haplotype structure in *long1* (around *Rosea*) is shown the upper four blocks, and the haplotype structure in *long* (around *Eluta*) is shown the lower four blocks. In each block, haplotypes are shown in rows in order of decreasing frequency. Each column corresponds to one of the 12 SNPs, the *Ros 1* (here SNP 3, 1 in *numerous*) and *El* (here SNP 17, in *numerous* 7) are fixed throughout the block and shown in red and yellow, respectively. Black tiles denote 0s and tiles in shades of grey denote 1s. The shade of grey corresponds to frequency of the haplotype in the group of plants in the block (the number of plants is shown in parentheses). Two horizontal dashed lines show the most frequent haplotypes accounting for first 50% and 90% haplotypes in the block, respectively.

CHAPTER 4

Image analysis methods

Although we often see studies focusing on simple or even discrete traits in studies of colouration the variation of "appearance" phenotypes found in nature is often more complex, continuous and high-dimensional.

When we started with this image analysis, the available phenotype information in this dataset consisted of three discrete manual scores provided by the photographer, describing three distinct features of the flowers: the amount of yellow pigment aurone (6 distinct gradual levels), the amount and on distribution pattern of the magenta pigment anthocyanin (9 distinct gradual levels) and the extent of magenta venation on the inner side of the upper petal. This comes with a number of problems. Firstly, the photographs have been scored by numerous people throughout the years, the scores are subjective, might not be consistent and are subject to human error. Secondly, this scoring system has been developed with Mendelian traits in mind [72] and while they may be suitable for their original purpose, the scope of the scores simply cannot describe *all* possible phenotypes.

Since each visible feature on a flower is only a part of a larger system that developed, is perceived and pollinated as a unit, we believe that we should strive to consider the phenotype of a flower in its entirety. Therefore, we believe, that when evaluating flower colour phenotype, we should use objective, automated phenotyping on the *entire images*.

Automated phenotyping is becoming a trend for plants, animal and for human samples, in academia and in industry (for example large-scale automated phenotyping of farmed plants with PlantEye by Phenospex [69]) alike. With large datasets available, complex questions in mind and large computational resources in hand, we decided to develop high-throughput automated pipeline especially suited for detecting fine differences in colouration of flowers in this thesis. To underline how timely our efforts are, there exist several other tools for automated measurements of plant colours from digital images published around the time this thesis was in preparation.

For example, in [35] the authors used digital camera images to quantify proportion of cover by and colours of flowers and leaves of various plants in the field and adjusted the measurements especially to the human perception.

In [20], the authors focus on accurate quantification of colour reflectance of flowers using digital cameras, summarising all specifics of the digital photography and helpful transformations of the measurements for achieving accurate colour measurements matching those from the

spectrophotometer rather than simply extracting raw numbers from photographs set to please the eye of a human consumer.

Perhaps the method most similar to ours [39] is a high-throughput colour phenotype measurement method using digital images. Interestingly, it includes analysis of *"shape-independent colour patterning by circular deformation"*, which could be well worth a try, although more suitable for images of apples featured in their manuscript than for images of complex flowers consisting of several highly variable (and developmentally relevant) parts.

Also, unlike the pipeline developed here, all methods found elsewhere seem to use simple thresholding for separating the object of interest from background (except [20], that does not mention segmentation at all), which proved ineffective in the case of highly variable *A. majus* flowers coming from the hybrid zone.

In this chapter, we describe the methods for image analysis used in chapter 5 and close with two experiments studying the effects on the photographic measurements caused by: firstly, the amount of pigment in a petal and secondly, the differences in lighting and photographic setup.

4.1 The photographs, the scores and the challenges

In the field we typically collect the data on plants that flower, or have flowered and hence potentially contributed to the next generation of plants. If present, a flower from such a plant is collected in a plastic tube with a screw-on lid and brought to the field station for imaging, together with leaf material that is later processed for genetic profiling. During the time between collection and photography the flower is mostly in the closed tube, in cool conditions, first in an insulation bag with an ice pack and later in a fridge. All flowers are photographed as soon as possible, but at most three days after collection. The green sepals (calyx) is removed immediately before imaging by a squeeze at the base, followed by a gentle pull from the petals.

The photographs have been taken with a standardised setup from summer 2015 on. The setup consists of a square black velvet "stage" to mount the flower on, a camera with fixed settings, on a tripod holding the camera in a fixed height above the stage and two lamps with classical, warm light lightbulbs. A typical photograph is shown in figure 4.1. Starting in 2021, a new "photobooth" setup is in use. The photobooth is enclosing the photographed area from five sides, leaving the front side open for manipulation with the flower, with camera lens facing the stage at the bottom through the opening in the top wall at a fixed distance, supported by a custom metal construction. The two lamps are replaced with day-light mimicking LED set attached to the upper wall of the photobooth. Since this light is of different intensity and less yellow, the camera settings are changed accordingly.

The photographic stage features the opening for mounting of the flower and the colour and light standardisation features (stripes), as well as space for identifying the flower source (usually by placing a lid with the plant ID sticker).

To capture as much information about the entire flower as possible, we typically take four images per flower. One is from the front with the "face" of the flower, tilted to show as much of the flower opening as possible. The second view is from the side, showing the floral tube and its length, as well as the profile of the corolla. The bottom view shows the inner side of the upper petals, offering typically the best view of the venation, if present. The last view is

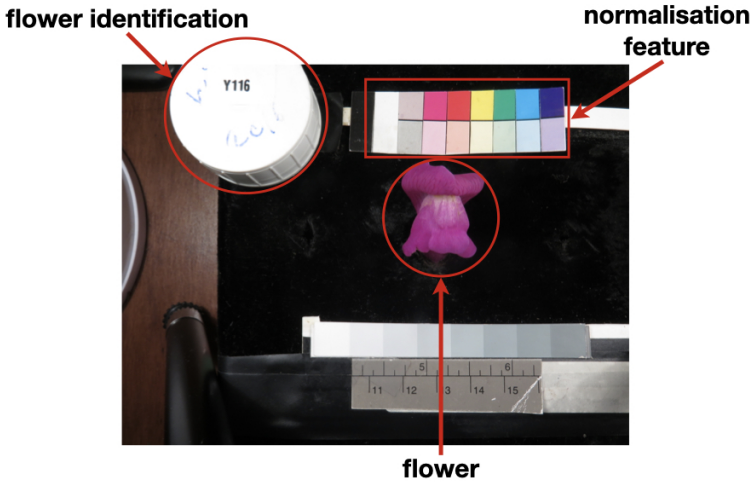


Figure 4.1: The stage is a black velvet box with a hole for mounting the flower, a colourful stripe used as a normalisation feature and a space for flower identification (here, a tube lid with Plant ID sticker).

from the top, showing the upper side of the tube and the exact spot where it meets the two upper petals.

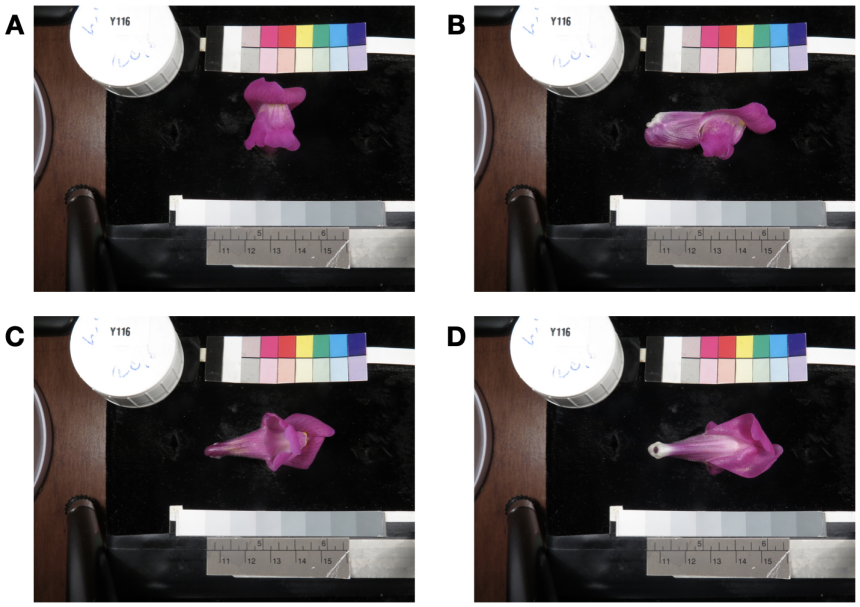


Figure 4.2: Each flower is photographed from four sides: from the front to show the "face" (A), from the side to show the colour pattern on the tube (B), from the bottom to show venation if any (C) and from the top (D).

So far, two approaches for quantifying phenotypes were used on our data. In the first approach, each flower image is examined at the time of taking the photograph and scored manually for the amount and distribution of yellow pigment (score 0.5 to 3), the amount and distribution of magenta pigment (score 0.5 to 5) and presence and structure of venation pattern (0.5-2), as described in figure 4.5 of [67], for table used for scoring in the field see figure 1.2. After

Figure 4.3: The manually selected discs: 1. upper petal (with visible venation), 2. top of the face (not always visible, usually shows some yellow pigmentation), 3. bottom of the face, 4. middle lower petal, 5. side lower petal. Credits: Taylor Reiter and David Field.

the field season the images with equal scores are grouped and manually checked for outliers in an attempt to avoid excessive variation in scores coming from different scorers. An advantage of such an approach is that the scores are readily available for all images. On the other hand, the crudeness of the scores does not cover all biologically relevant variation.

In the second approach, developed and carried out by David Field and Taylor Reiter, images are normalised for lighting using a black and white normalisation feature present in each image, and RGB measurements are taken from small standardised discs manually placed in specific parts of the flower. From these, other variables can be calculated, such as hue, saturation, distribution and variation of the pigment in each standardised circle. While this approach helps to uncover possibly important differences by taking measurements from biologically relevant parts of the flower, the manual placing of the circles makes it less reproducible and not scalable to the extent of the whole dataset. Furthermore, using the normalisation feature directly might be harmful, as shown later in this chapter.

The main difference from what has been done so far is that our methods are fully automatic and hence suitable for large dataset without extensive manual labour. The strength of this approach is especially in reproducibility, which typically suffers from human input, especially when several different individuals contribute to the same project, possibly long times apart.

In short, our method takes images of flowers together with permanent normalisation features and first extracts the flowers from the background. The normalisation process then identifies the permanent feature in each image and produces a list of values characterising the lighting conditions for each image. The final result is a list of values characterising the colour of flowers in the images. To achieve this, we combined existing image analysis software with scripts exploiting Python image libraries and bash to create a flexible, scalable and user-friendly pipeline, compatible with use of a cluster.

The process of automatic phenotype quantification is challenging, as the flowers come in a range of different colours, shapes, sizes and conditions, from plump fresh flowers to wilted and partially destroyed flowers and flowers in various stages of decay and the pipeline needs to be able to distinguish the flower pixels regardless of all the other variables. Therefore, the pipeline has been trained and tested on flowers with different phenotypes, as well as damaged flowers and post-processing steps have been added to make it as robust as possible.

Also, although a lot of effort has been invested in keeping the keeping the photographic conditions such as lighting and the position of the stage as similar as possible, it is not possible to keep them constant with two flexible lamps, differing light conditions in the field station and more than six photographers imaging and scoring flowers throughout the years. Therefore, we developed a normalisation system resulting in a set of indicators characterising the light conditions and the geometry of the setup that can be easily calculated for each flower image using automatic recognition of the normalisation feature.

Another feature of our approach is its unbiasedness. By using all the flower pixels at once we avoid the bias that would be introduced by human choice of features. However, there will still be bias connected to acquisition of the data which cannot be avoided. This will include choice of flowers, positioning of the flowers on the photography stage and sensitivity of the camera chip. While the first biases may be obvious, the sensitivity of the camera chip is especially an interesting one, as it is different from both the sensitivity of the pollinator's eye (more on this topic in section 6.2.3) and pigment combination and flower formation. Therefore, all of this should be taken into account when interpreting the results.

At the same time, the incorporation of spatial information into the analysis may become challenging. Distribution of the pigment in a flower has been shown to have an effect on pollinator attraction and success [74, 55, 28]. Therefore, while analysing a "soup of pixels", may be attractive, due to unbiasedness and simplicity, it will necessarily lack biologically relevant information and therefore should be taken with caution and ideally combined with information on flower parts location in the future.

Please note that pipelines such as this one can be updated and improved forever. Therefore, in addition to methods implemented and used in the other chapters we provide possible changes and improvements *in italics* throughout, as well as ideas for the future directions (especially automatic flower part identification) at the end of this chapter.

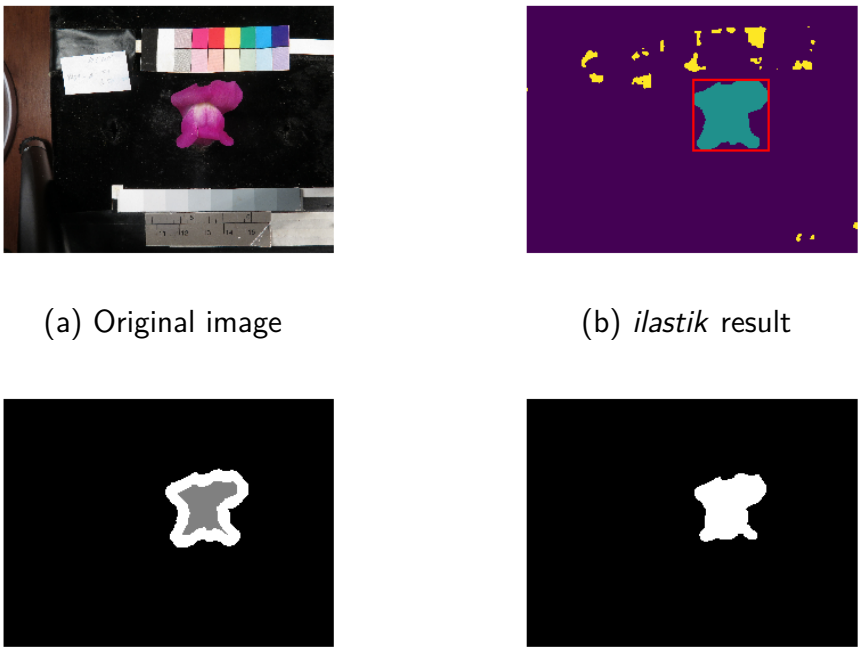
4.2 Identifying the flower pixels in images

The first step in the analysis is separating the flower pixels from background in the image. The result would be a mask, i.e. a matrix of zeroes and ones with the same size as is the original image, with ones at position of the flower pixels and zeroes elsewhere. To accomplish this, we used the image analysis software *ilastik* [5] and a set of Python image analysis libraries together with my own Python and bash scripts which combined to a flexible and user friendly pipeline, compatible with use of cluster.

4.2.1 Finding the flowers with *ilastik*

First, we created *ilastik* pixel classification and an object prediction project to find the flowers in raw images. The project can then be used from terminal, either on a local machine or remotely, to process large numbers of raw images.

First, the *ilastik* project needs to be manually "trained" to recognise flower pixels from the background pixels based on pixel properties and user annotations. Using Random Forest [27] to select relevant qualities of pixels and their surroundings (colour, smoothness, etc.) and assign weights to them, this step results in assigning a probability of being inside a flower object for each individual pixel and serves as a basis for object classification.



(a) Original image (b) *ilastik* result

(c) Mask for grab-cut (d) Output mask

Figure 4.4: The stages of image analysis. Starting with the original image (a), first, the pixels are labelled by *ilastik* to distinguish the most likely flower pixels (b) from the background. Then, a safety margin is added on the inside and outside to the most likely flower pixels to mark the pixels that are for sure inside of the flower (grey) and outside of the flower (black), respectively (c) and this altered mask is given to the grab-cut algorithm. The end result is a clean mask indicating where there were flower pixels in the original photograph (d).

In the next step, the pixel-wise probability map is thresholded to provide segmentation of the raw images and the individual objects are classified based on a subset of object properties and user input. For each input image this step results in one new image containing the mask and a .csv file detailing the result.

We train the *ilastik* pixel classification and object prediction projects on 10-15 images of flowers with highly varying phenotypes. We run the trained project from terminal or cluster in "headless" mode to batch process large numbers of similar images (for details see 4.8). An example result of such pixel classification is in figure 4.4 (b).

4.2.2 Fine-tuning of the masks using Python functions

Sometimes additional adjustments of the masks are needed. To perform these we use simple morphological transformations and the algorithm grab-cut [53], implemented in Python libraries *OpenCV* [8] and *scikit-image* [70].

We use simple morphological transformations such as erosion and dilation from *OpenCV*. Combinations of the two (also known as opening and closing) is widely used to remove noise. They do this first by "eroding", i.e. shaving off from the edges of all objects in the image using a specified kernel, thus annihilating minuscule objects which were likely unimportant: either noise, or perhaps dust, small insects or pollen grains. This is followed by "dilating" the

remaining large and most likely significant objects by inflating the edges to add the shavings back and restore their original shape.

Since this usually results in poor quality of masks and we would not want to miss any flower pixels, we use a combination of these transforms with the grab-cut algorithm [53] to perfect the rugged margins of the masks. The grab-cut algorithm was designed to efficiently separate image foreground from the background. It fits Gaussian Mixture Models over the user-specified foreground and background pixels, then clusters them into these two categories based on colour similarity and proximity, all of this being automated and reproducible.

To create the foreground and background models for the grab-cut algorithm, we exploit the erosion and dilation algorithms on the imperfect, but appropriately placed masks from *ilastik*. First, we erode the *ilastik* mask to shave off the uncertain edges and include nothing but high quality flower pixels and label these pixels as foreground. Pixels from outside of the foreground which belong to the (very generous) dilation of the original mask to include all flower pixels and some background we label probable foreground and the rest is labelled as background. We use this upgraded mask (in figure 4.4 (c)) as an input for the grab-cut algorithm, which identifies the foreground and background on the basis of these labels and proximity.

To yield the final mask (in figure 4.4 (d)), the result is cleaned from noise by erosion and dilation and the object at the flower position is selected. Finally, this mask for each flower image is saved for further analysis, as well as for the future use.

Given the paths to the images, this procedure is fully automated, without need for specifying any parameters. However, it sometimes fails due to poor quality of the image or due to the flower being shown from a different angle than expected (if an incorrect image path was given to the pipeline). Very rarely, it can fail for different reasons, but this is so rare, that we do not have an estimate for accuracy of this procedure. However, since simple thresholding used for production of masks in other published methods is a subset of what *ilastik* does, we expect this to perform much better than any of them (this, of course, can be tested).

4.3 Normalising light and colour using permanent features

Although a lot of care was put to making sure that the images in our dataset were taken in constant conditions, they were still taken by several people, in several locations, throughout several years. However, there are several permanent features on the photographic stage to be able to control for these conditions afterwards. Here, we show how we used the permanent colour normalisation feature to extract information on lighting information and information on distance of the stage from the camera lens for the purpose of light, colour and size normalisation of the images.

First, we identified position of the permanent feature using template matching function from *OpenCV* library similar to figure 4.5. In order to do this, we created a template by cutting out a typical permanent feature from a well-lit image with normal distance between camera and the lens (in this step, it would also be possible to use a mean of such cutouts instead). The template matching function then slides the template through all possible positions in the image (without rotating it) and outputs a matrix with values describing the fit of the template to the image at each position. The best fit can then be found by finding the position of maximum in this matrix.

Figure 4.5: The template to be matched to a flower image and its most likely position right at the permanent colour normalisation stripe.

After we identify the the most likely position, we only work with this section from now on and save it for possible further analyses. Inside the section, we identify areas (sets of pixels) most likely to be magenta, red, yellow, green, cyan and blue. For each colour, we do this by defining a function $a \cdot \Delta hue + b \cdot \Delta saturation$ defining a kind of distance from the desired hue and saturation, weighted by a and b, with $a \gg b$, because the differences in hue are much more important than those in saturation, but once the hue matches, the saturation may still be relevant. We tested various combinations of a and b on a series of differently lit images and a ratio of $\frac{a}{b} \approx 15$ works well for all of them. Since the permanent feature consists of rectangles of the same colour, the function values on all section pixels ordered by output will appear approximately step-wise. To identify the best fitting pixels we then need to identify position of the first "step" and the pixels that lie before it.

This whole procedure could be replaced by segmentation with k-means clustering by colour, since we know exactly how many clusters of similarly coloured pixels there should be. However, this would approach could run into problems with very bright images where too many rectangle areas appear white.

For each of the colour pixel sets, we find mean and variance, and quantiles 0 (minimum), 0.25 (first quartile), 0.5 (median), 0.75 (third quartile) and 1 (maximum) of hue, saturation and value. We also find median of their coordinates and fit a simple linear regression through all complete and typically well-identified ones (red, magenta, green and cyan, blue is typically incomplete and yellow poorly identified in bright images) to estimate the most likely width of a single rectangle. We save the value of most likely width of a colour rectangle as a size unit and the R^2 of the regression as quality of fit for future reference.

If the y-coordinates have either an increasing or a decreasing tendency, this value could be adjusted by the estimated angle to account for possible horizontal rotation of the stage relative to the camera lens.

All of these numeric values are then stored in a spreadsheet for future reference, together with plant ID, image name and paths to relevant image files, with values for one flower image corresponding to one row.

4.4 The image statistics

Once we possess a good quality mask identifying the flower pixels in the image, we would like to extract flower characteristics on pigment hue, concentration and distribution, as well as details about flower size and shape. To do this, we use thresholds on hue and saturation to select magenta, orange, yellow, wide-sense magenta, wide-sense yellow (larger scale of hues, explanation below) pink, white and "other" pixels. Then we count the number of these pixels

and the proportion they contribute to the number of pixels in the whole flower, as well as mean, median, variance, minimum, maximum, first and third quartile of their hue, saturation, value and red, green and blue values.

The scale of hue starts in red with $hue = 0$, grows through orange, yellow, green, cyan, blue and magenta and ends in red with $hue = 1$ again. This is very impractical for us, as we work on a scale from magenta through to yellow and while we would like to see this as a smooth scale, there would be a step from 1 to 0 at read if we used the normal hue. Therefore, we use transformed Hue with a capital H such that: $Hue = hue - round(hue)$ which starts at -0.5 at cyan, grows through blue, magenta, red ($hue \approx 0$), orange, yellow and green and ends in cyan again at $hue = 0.5$. This is particularly useful, as no part of the flowers comes close to the cyan hue and we can then calculate shifts in hue easily.

4.4.1 The aurone and the anthocyanin pigments

The ultimate goal of this analysis is to determine the presence of yellow aurone and, in particular, magenta anthocyanine, their concentration and distribution. Patches with pure magenta or pure yellow pigments would be easy to single out, as their hues differ significantly. However, they are often present in the same spot within the flower, making this spot appear red, or orange, depending on concentrations of the pigments. Figuring out ratios of the pigments based on the resulting appearance in the photograph is a complex issue which we do not aspire to solve here. Instead, we use two special colour categories for the pixels: wide-sense magenta and wide-sense yellow categories, to include all pixels that have some magenta pigment in them (blue-purple through magenta and red, all the way to orange-yellow) and those that likely contain yellow pigment in them (red through orange and yellow all the way to lime green). This way, we can asses the approximate distribution of both hues across the flower, as well as tell whether the pigments coincide in the same areas of the flower without making too many assumptions.

In particular, we do not want to assume that resulting colour is a simple linear combination of the two pigment amounts, as this may be much more complex. However, a simple linear model like this can be tested.

For illustration, we show histograms of transformed hues and saturations of wide-sense and narrow-sense of magenta pixels from five phenotypically different flowers in figure 4.6. One can see that although flower 1 and flower 2 would both be classified as full magenta, the magenta hue of flower 1 is shifted more toward the more positive values, compared to flower 2. In other words, flower 1 appears more red. We can test the concept of narrow- and wide-sense magenta on case of flower 4 with magenta venation and pale anthocyanine coloration mixed with yellow, visible on the upper petal. This means it has some wide-sense magenta pixels and a few of these, not very saturated ones even fall into the narrow-sense magenta category. We can compare this to flower 3, which has some pixels classified in the generous wide-sense magenta, but no narrow-sense magenta pixels. Looking at flower 5 once can see that the analysis is works for flowers of low quality too, it just has a peak in an unusual hue range, between red and yellow, corresponding to the brown, damaged parts of the flower.

4.5 The spectrophotometry experiment

The goal of this experiment is to validate the method of quantifying pigment concentrations from photographs and describe a relationship between the spectrophotometric and photographic

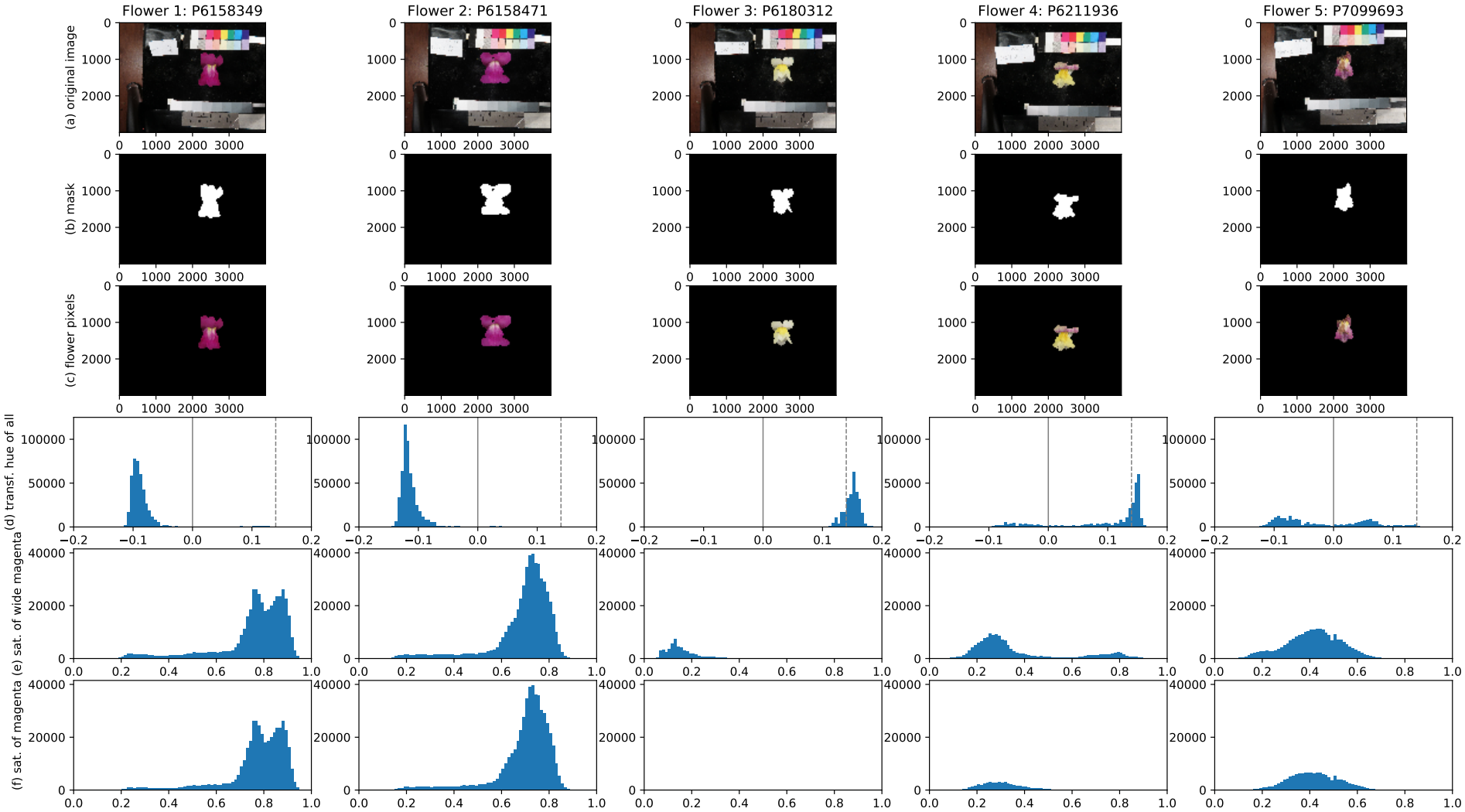


Figure 4.6: Image statistics for five flowers with different phenotypes (a), their masks (b) all analysed flower pixels (c) histograms of transformed hue for all flower pixels (d) and saturation of both wide-sense and narrow-sense magenta pixels (e and f). The arbitrary threshold for wide- and narrow-sense magenta hue are indicated in (d) by dashed and solid grey vertical lines, respectively.

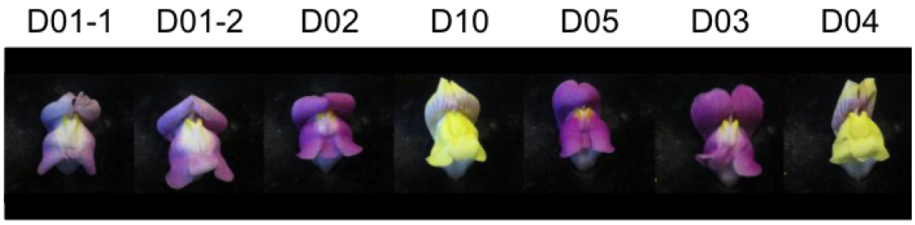


Figure 4.7: Images of flowers used in this experiment with IDs, D01-1 and D01-2 come from the same plant D01.

measurements.

In spectrophotometry, transmittance can be calculated as $T = \frac{I_t}{I_0}$, where I_0 is the intensity of light of a certain wavelength shone on the sample and I_t is the transmitted, "leftover" light intensity. The spectrophotometer machine typical output is absorbance is $A = -log_{10}T$. This measure is widely used to compare concentrations of pigment c in solutions, since, absorbance is directly proportional to pigment concentration in the sample according to Beer-Lambert's law: $A = \epsilon lc$, where l is the length of the trajectory the light had to travel through the sample.

We selected seven flowers from six plants (one flower from two yellow, three magenta plants and two flowers from one pink plant, see figure 4.7) and photographed these flowers as usual. We cut out six same sized discs out of each flower, two from upper petals, two from the face and two from lower lower petals, always in symmetrical left and right pairs We photographed them and then put each of the discs in a separate Eppendorf tube with 500 mL of solvent (hydrochloric acid diluted in methanol). After two hours 300 mL of the flower petal extract was transferred from each Eppendorf tube to a separate well of the 96-well plate. One row of wells was filled with the same amount of plain solvent. Using Spectrophotometer Biotek Synergy H1 platereader (I04.EG.017) the absorbances at 400, 496 and 512 nm were read at each of these wells as well as of some empty wells and some wells with pure solvent for comparison.

Since absorbance is directly proportional both to concentration c and to the length of trajectory the light travels through the sample l, in our context, there are two inherent sources of error stemming from pipetting. First, the amount of liquid in the Eppendorf tubes: c depends on the *amount of solvent* as well as amount of the pigment. Second, when transferring the same extract into the 96-well plates with flat bottom, l is directly proportional to the amount of flower petal extract in a well.

To read the colour intensities from the disc photographs we separated the discs from the background using an Ilastik classifier and a procedure similar to that described in section 4.2.1, but trained on images of cutout flower petal discs on the same background. We extracted the hue, saturation and value (HSV) values from all pixels coming from the discs rather than from the background. For each disc, we subset the pixels and created two sets of pixels with hues in yellow or magenta ranges, respectively. Then we calculated *mean yellow and magenta saturations* of the discs summing up the saturations at all pixels from yellow and magenta sets respectively, and normalising by the *number of pixels detected in each disc*. We decided to use the *number of pixels detected in each disc* for normalisation, rather than normalise by average area or not normalising at all, because some of the discs curl up slightly and not all of the pixels for all the discs are visible in the photographs. Therefore, calculated this way, the *mean saturations* should better reflect the overall amount of the pigments in the discs and

correspond to concentrations measured via spectrophotometry.

To check the whole absorbance spectra for any additional peaks, we prepared two additional tubes with extract, one containing a mixture of *A. m. striatum* flowers and the other one with *A. m. pseudomajus* flowers. Allowing sufficient time for the pigments to dissolve, several 300 mL samples of the flower petal extracts were transferred from either of the tubes to a separate well of the 96-well plate. Using the same machine, the whole spectra were read at 5 nm intervals for all wells containing the extracts, as well as some wells with 300 mL of pure solvent for comparison.

4.5.1 Results

As expected, the magenta flowers have much lower concentration of yellow pigment than the yellow flowers and vice versa (Fig. 4.8). The left-right symmetric replicates seem to be more similar to each other in measurements at 512 nm than at the non-zero measurements at 400nm, which can be explained by localisation of the pigment; the distribution of yellow pigment in the petal (especially in dark yellow flowers) is quite patchy and cutting the two discs exactly symmetrically can be challenging.

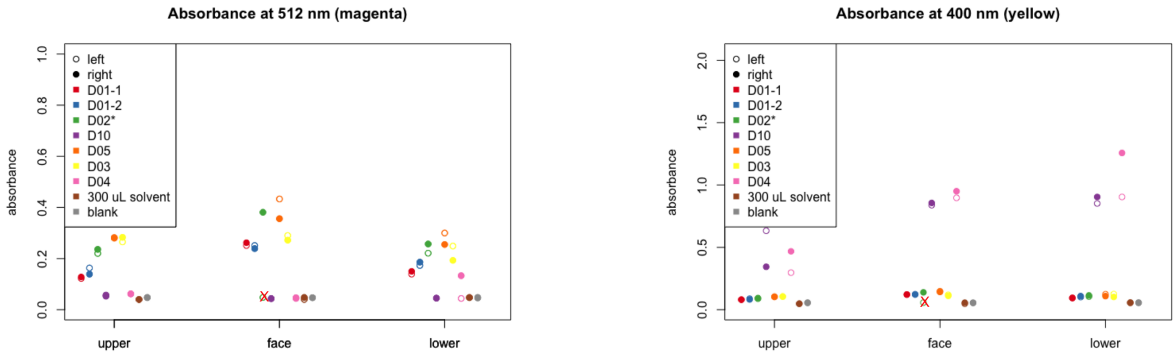


Figure 4.8: Absorbances at 512 nm and 400 nm for discs taken from upper petals from flowers D01-1, D01-2, , face part and from the lower petals, in symmetric left and right couples, empty wells and wells with pure solvent. Red cross denotes the missing extract of the left face disc from plant D02.

The normalised saturation both for magenta and for yellow pixels in figure 4.9 seem to be biased towards higher values in the discs taken from the right side of the flower, which can be simply explained by their higher distance from the lamp and thus being less bleached by the light in the photographs. And indeed, the difference is supported by data, using one-sided paired t-tests resulting in p-values of 0.02024 and 0.01067 for magenta and yellow, respectively (the near-zero measurements in irrelevant flowers were excluded).

Comparing the absorbances with the mean saturations it seems that these two measures indeed correlate (for both comparison the Pearson's correlation coefficient is above 0.9) and that the photographs indeed carry a valuable information about the amount of the anthocyanine and aurone pigments in flower petals. However, there are several factors which can obscure this relationship. Firstly, position of the pigment within the tissue of the petal: the photographic measurements of mean saturation will differ based on the depth of the pigment, although usually, the pigment is present in the top layer of the cells (do these have a name?). Secondly, the lighting can make a big difference in the readings of the mean saturations, as it can shift

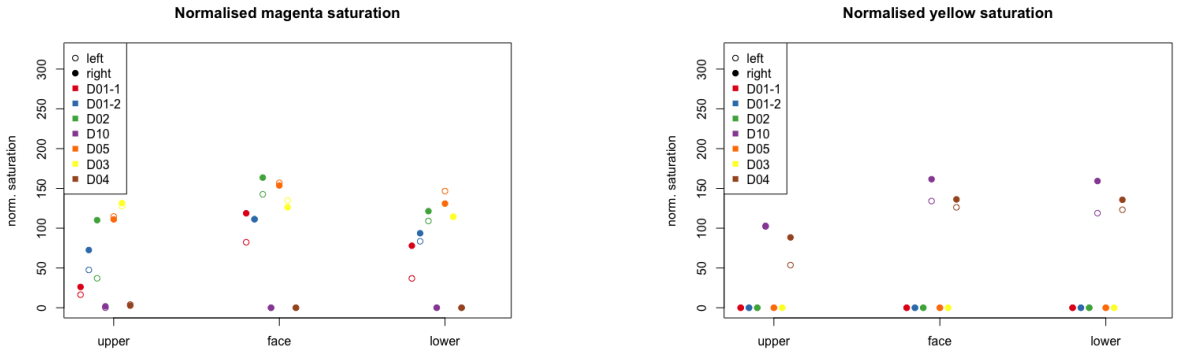


Figure 4.9: Normalised saturations of pixels classified as magenta or yellow for discs taken from upper petals, face part and from the lower petals from flowers D01-1, D01-2, D10, etc., in symmetric left and right couples. There seems to be a systematic bias for higher values in the discs taken from the right side of the flower.

all hues depending on it's colour, shift the hues in shadows of 3-dimensional objects towards blue and even "bleach" the saturation of the objects that are closer as seen in figure 4.9.

To adjust for the effect of position relative to the lamp (i.e. the discs coming from left or right side of the flower), we decided to make a linear regression:

$$\textit{normalised saturation} \sim \textit{absorbance} + position + \textit{flower part},$$

where *flower part* is either upper petal, face or lower petal. We made two separate regressions, one for yellow saturations and absorbances with measurements from yellow flowers only and one for magenta saturations and absorbances with measurements from magenta (and pink) flowers, as there would be too many near-zero values confounding the results otherwise.

The results of the linear regression in table 4.1) show that in magenta flowers, the absorbance explains most of the variation in normalised saturation of pixels in magenta range. The discs taken from the lower petals are significantly more saturated than those from the face (or upper petals) and the effect is both about as significant and as large as the effect of being more distanced from the lamp. Since both of the yellow flowers were very full yellow, the simple factor of coming from a flower categorised as "yellow" in this regression is enough to result in a high saturation of a disc (hence the significant Intercept and uninformative absorbance in the yellow range). Consistent with empirical observations, there is a trend of discs coming from upper petals (and the discs more distant from the lamp) being more saturated in yellow flowers, but more measurements on more variable yellow flowers would be needed to confirm this.

Whole spectrum absorbance

The lines resulting from the whole spectra reading can be assigned to three very different, easily recognisable clusters: one for yellow and for magenta flower petal extracts and one for the solvent. Within the clusters, the resulting lines are parallel. Given the same concentrations of extracts in each well within the cluster, this points to very accurate measurements and possible differences in amounts of the extracts.

In agreement with available literature, the very distinct absorbance peak for magenta (more specifically for *antirrhinin*) is at 512 nm [34]. However, this peak is rather wide and although

Table 4.1: Linear Regression Results

	Dependent variable:	
	Magenta norm. sat.	Yellow norm. sat.
	(1)	(2)
magenta_absorbance	481.194*** (51.568)	
yellow_absorbance		17.701 (40.207)
is_right	16.425** (6.182)	19.925* (9.840)
partlower	19.102** (8.986)	−6.935 (12.284)
partupper	−4.047 (8.999)	−44.795* (21.532)
Intercept	−26.603 (17.055)	113.881** (35.693)
Observations	29	12
R^2	0.849	0.810
Adjusted R^2	0.823	0.701
Note:		*p<0.1; **p<0.05; ***p<0.01

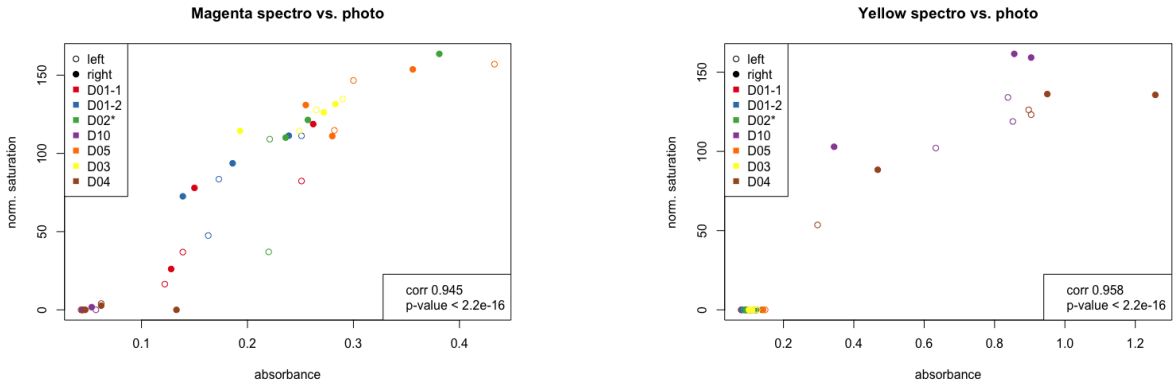


Figure 4.10: Comparison of absorbances measured at 512 nm and 400nm and mean (normalised) saturations of pixels in magenta and yellow hue range for all flower discs.

not perfect, measurements at 512 nm are a reasonably good measure of the amount of anthocyanines in the flower petal tissue.

Both yellow and magenta extracts have a distinctive yellow absorbance peak around 390 - 400 nm typical for aurone, as expected.

What remains baffling is the high absorbance at lower wavelengths, slightly obscuring the yellow peak in readings of both yellow and magenta flower petal extracts. It has been suggested that this is due to high concentration of flavones with absorbance peaks in lower wavelengths [34], which seems to be plausible, as these extreme values only occur in measurements of flower petal extracts and not in those of the pure solvent.

The results of this experiment suggest that it is reasonable to use photographs to estimate the concentrations of aurone and anthocyanine in flower petal tissues. However, lighting can have an effect on both saturation and hue of the resulting images and this should be taken into consideration both when producing the images and when interpreting the results. Where possible, these effects should be adjusted for.

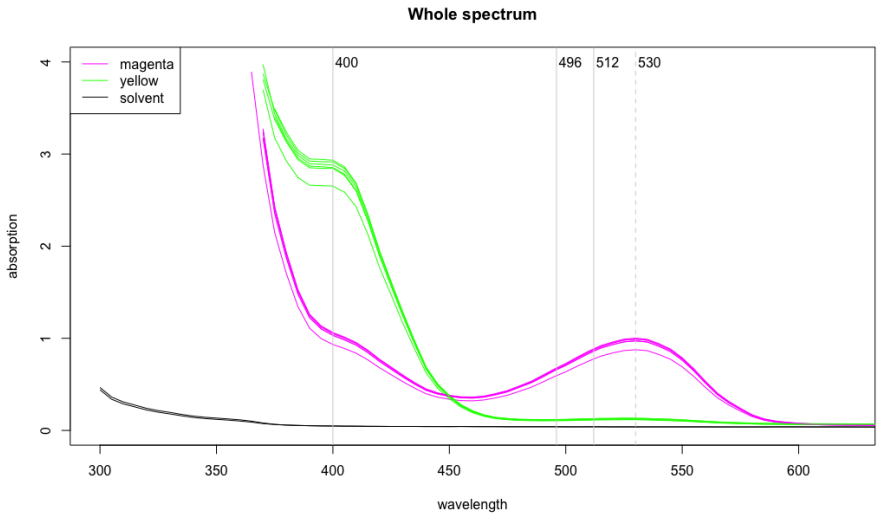


Figure 4.11: Whole spectrum absorbance measurements of magenta and yellow flower petal extracts and pure solvent. Multiple lines show replicate measurements of the same three solutions.

4.6 The setup experiment

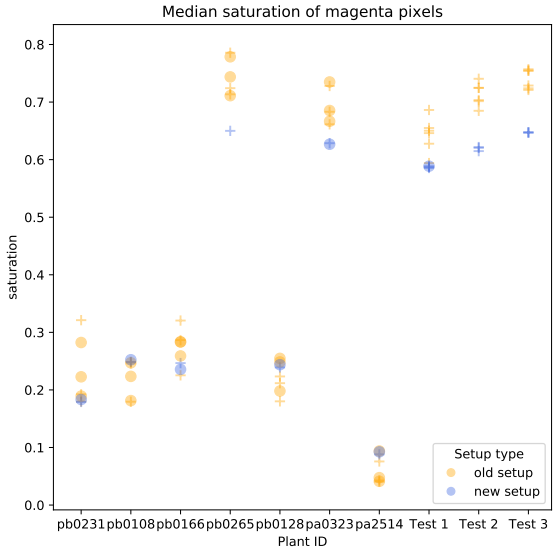
There is a complex relationship between lighting and colour values of a three-dimensional object using in a photograph and a flat permanent colour normalisation feature. Even more so, if they are in different positions relative to the source of light and consist of such different materials, as a flower of *A. majus* and a glossy colour paper stripe.

Therefore, we decided to investigate this issue (and test the new setup) by comparing images of the same flowers, taken under varying circumstances. In particular, we are comparing the "old setup" with two lamps and "warm" light producing light bulbs to the "new setup", with a fixed led source of diffused white light. We also varied the distance from the stage to the camera lens and the lamp positions in the old setup to mimic the accidental alterations the old setup is particularly prone to and demonstrably experienced throughout the years. This dataset was produced by Arka Pal.

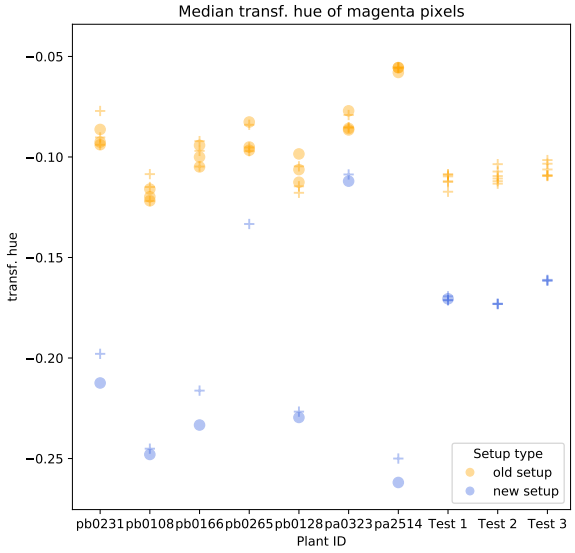
First, the results in figure 4.12 suggest that the position of the photo stage relative to camera lens, as well as the position of the lamps has an effect both on saturation and on hue measured in the flower images, therefore we recommend using a physically fixed setup with as little variance as possible.

Second, figure 4.13 shows that while in the new setup the saturation and hue of the permanent feature are quite stable, in the old setup they vary with the position of the lamps with no relation to the hue and saturation measured in the corresponding flower images. Therefore, we do not recommend using the data from normalisation part of the pipeline directly for normalising of the values. However, as shown in figure 4.14 they can still help by providing information on the photographic setup in the downstream statistical analysis, for example, as a factor grouping images with a similar setup and lighting conditions together.

Figure 4.12: Comparison of images of the given flowers taken with old and new setup (orange and blue, respectively), with low (○) or high (+) stage position.



(a) The highly saturated flowers seem to appear more saturated in the old setup, but there is no significant trend in less saturated flowers.

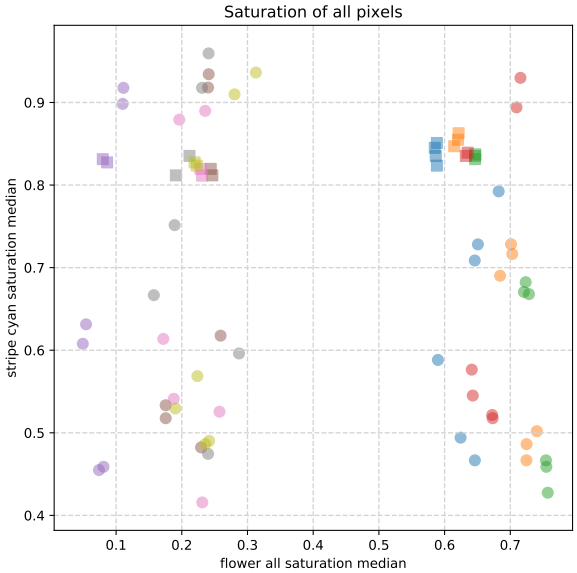


(b) As expected, the hue of the magenta pixels appears more blue in the cold white light of new setup.

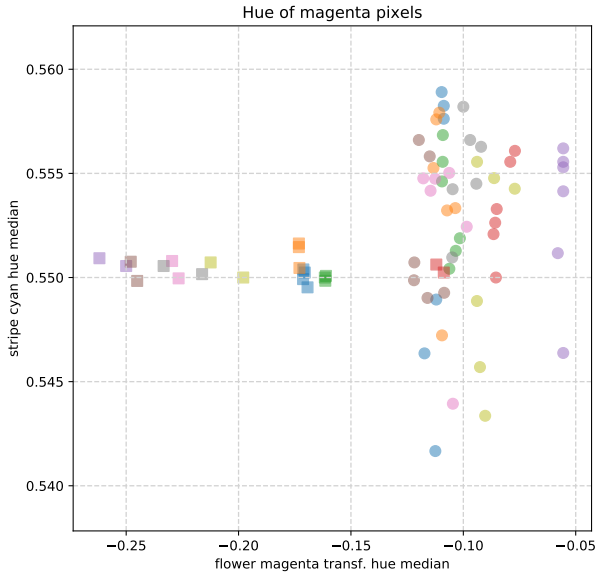
4.7 Summary and future directions

In this chapter, we described the image processing pipeline. We also showed that the data from normalisation should be used for categorising the photographic setup rather than directly for normalisation. This is due to the fact that the flower has a different placement of the pigment, as well three-dimensional structure protruding from the smooth background, and

Figure 4.13: Comparison of flower vs. permanent feature appearance within images taken in old ("○") and new ("□") setup, each colour stands for one flower, each with a different



(a) The flowers clearly separate to more and less saturated. Overall, the differences in saturation seem to be smaller in both normalisation stripe and in flowers for the new setup. While the saturation of the stripe varies wildly in the old setup, depending on the position of the lamps, the variation in



(b) While the permanent normalisation stripe hue is largely constant in the new setup, it varies wildly in the old setup, dependent on the position of the lamps. Interestingly, in the old setup the flower hue appearance varies about as much as the stripe hue appearance with the position of the lamps (≈ 0.015 hue units), but the hues are not correlated (for images of the same flowers).

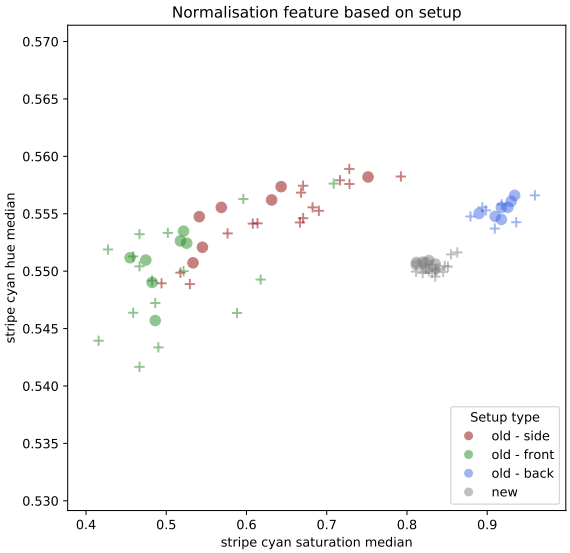


Figure 4.14: The values measured in the permanent normalisation feature based on setup (colour-coded), low (○) or high (+) stage position.

thus it is illuminated (and reflects light) differently than the flat normalisation feature, some centimetres above.

This approach has limitations, in particular it does not explicitly address the spatial distribution of the pigment throughout the flower. However, spatial distribution of the pigment in a flower has an effect on pollinator attraction and success as implied in [74, 55, 28].Therefore it would be beneficial to include spatial information on the distribution of pigments throughout the flower itself. This is possible by manually annotating biologically relevant flower parts or landmarks for each raw image or by automation of this process. To keep the approach scalable for large amounts of raw images, unbiased by human interference and fully reproducible, we suggest using existing deep learning tools to identify the whole flower parts.

The process of constructing a successful deep learning tool for identifying the flower parts starts with producing a training set. The training set should consist of hundreds of raw images together with their corresponding masks outlining individual flower parts, for example to "face", upper and lower petals and the rest, as those proposed in figure 4.15. These parts of flowers are a reasonable choice not only because they likely develop and are mostly coloured together as units, but also because they typically have the clearest borders. Also, this compartmentalisation could help to automatically distinguish relevant information on the floral guides, in particular on the yellow patch in the face, or venation on the upper petal. Such masks can be produced manually, using Photoshop. For identifying the flower parts, a machine learning algorithm Mask R-CNN [26] trained on such dataset can be used to extract defined areas corresponding to individual flower parts for each original image.

4.8 Detailed instructions

Complete workflow, names and descriptions of scripts, required software versions, folder structures and file formats.

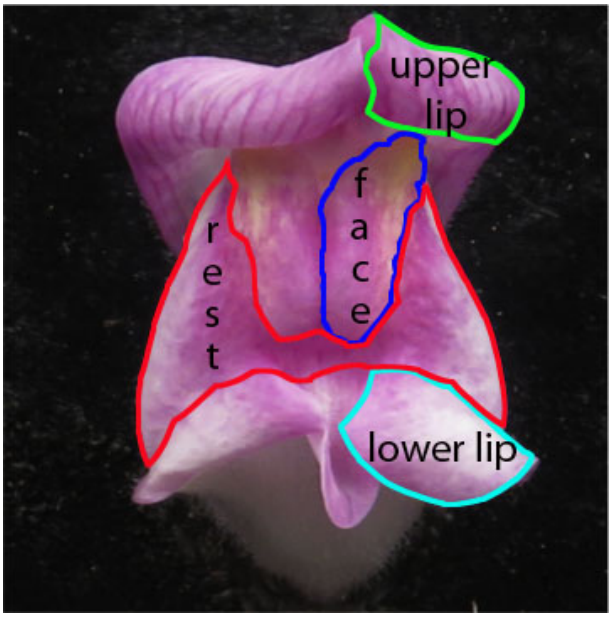


Figure 4.15: Proposed flower parts separation, with petals referred to as "lips". These parts of flowers are a reasonable choice not only because they likely develop and are coloured together as units, but also because they typically have the clearest borders and because they could help to distinguish information on floral guides.

First, we need to identify all images we want to process and find their paths from within the archive. To do this, we use Python script of type "produce_paths" with an input of Plant IDs we want the flowers from. The script will attempt to read the year from the Plant ID letter and will automatically look into the field data, but both the photoscores file and the raw photographs folder can be overwritten instead. The "produce_paths" script results in a .csv file with plant ID, year, photo code (in form Pmddnnnn), or a short message in case something went wrong (image missing, in a wrong folder, etc.). Sometimes, the scorer forgot to update the date when generating the photo codes and the image is in a folder with different (later) date. The script does look for images matching the description later, but its options here are limited and image still can exist somewhere in some wrong folder even if the script cannot find it. We copy the face images to our working directory using Python script "copy_faces". Next, we need to downsize the relevant images using Python script "down", resulting in a folder of .png images with suffix "_10th.png".

To use *ilastik* and produce the first predictions, we need to produce bash scripts that will run *ilastik* on batches of several images. To produce the scripts we use a Python script that produces folder for the scripts and then produces the bash scripts calling ilastik to classify the images using an ilastik project (.ilp). When running these bash scripts one needs to be careful, as the project can only be used by one bash script at a time. Therefore, we recommend to use a bash script "multi_job.sh" that will run the all scripts in specified bash scripts folder one after another, avoiding multiple jobs attempting using the same *ilastik* project. If in need of faster processing, the *ilastik* project can be copied over multiple times (with different names) and scripts can be generated so that the scripts running in parallel use different copies of the *ilastik* project.

Once the *ilastik* results are available, they can be used to produce final masks using Python script "make_masks". Using the final masks, we can proceed to calculate image statistics

using Python script "image_stats", resulting in a .csv file with Plant ID, year, photo code and a set of image statistics discussed earlier in the chapter, or an error message.

If we wish to calculate the normalisation characteristics of our images, we run the Python script "normfeat" on the downsized face images. This will result in a .csv file with normalisation data (or an error message) for each image. Quite importantly, this script needs to load the normalisation stripe template image "normfeat_10th.png" to match to the downsized images.

All Python and bash script templates and other necessary files and a couple of example projects (*cline* for phenotypic clines, *cali* for calibration experiment, and *setup* for testing the old and the new setup) are available in the public IST repository coupled with this thesis.

CHAPTER 5

Phenotypic clines across the hybrid zone

A hybrid zone is an area with increased frequency of hybrids in areas where two distinct parental populations came into contact. Usually, one can observe *gradients* in frequencies of traits typical for the parental populations perpendicular to the direction of the contact. These gradients, also known as *clines* are maintained by selection despite the dispersal, with steep, narrow clines marking areas where the two parental populations are kept strictly isolated, whereas wider clines revealing areas of more mixing. Therefore, besides maintenance of a cline being a proof of selection's existence, the steepness of a cline is a measure of selection strength. Consequentially, clines have been thoroughly studied in many systems all around the world, including toads (genus *Bombina*) in Croatia, butterflies (genus *Heliconia*) in several countries of Central and South America and monkeyflowers (genus *Mimulus*) in the United States.

In this chapter we study a hybrid zone near Planoles in Spain connecting two parental populations of snapdragon plants *Antirrhinum majus* differing by little else than the colour of their flowers: the magenta flowers of subspecies *A. m. pseudomajus* coming from the east and the yellow subspecies *A. m. striatum* stretching on the west from the hybrid zone. (Although these two subspecies come into contact in several places in France and Spain, this hybrid zone is especially rich in hybrids, often distinct in colour.)

We use the *multivariate* colour measurements automatically derived from *entire* front ("face") flower images to study the relationship between flower colour of *Antirrhinum* flowers and geography within the hybrid zone in Planoles. In particular, we describe a *phenotypic cline* passing through the hybrid zone in order to understand the effects of phenotypic differences on maintaining the hybrid zone and to compare it to previously described genotypic cline [68]. In order to do this, we use subset of $5,000$ plants sampled between 2016 and 2019, enriched in individuals growing in the hybrid zone core.

In general, we classify phenotypes into six categories based on manual red (anthocyanin) and yellow (aurone) scores (for table used for scoring in the field see figure 1.2), as described in table 5.1, which we are going to use to train classifiers, test our results and for visualisation. These categories are defined based on division across two axes, regulated by two distinct, unlinked sets of genes (plants dominant and recessive for locus *Locus* shown as *LOC* and *loc*, respectively).

The first axis stretches along the yellow manual score, and describes the amount of aurone. It is mostly governed by *Sulfurea* and *Flavia*. Along the yellow axis, we divide the plants

			Yellow score	
			pale (< 1.5)	full (≥ 1.5)
			SULF	sulf
Red score	pale (< 1.5)	ros	white	yellow
	patchy (between 1.5 and 2.5)	ROSEL	pink	weak orange
	full (≥ 3)	ROSel	magenta	dark orange

Table 5.1: Six phenotypic categories based on manual scores and typical *Sulfurea*, *Rosea* and *Eluta* genotypes. Plants dominant and recessive for locus *Locus* are shown as *LOC* and *loc*, respectively [68, 7].

to "pale" yellow (mostly just a yellow patch near the opening of the flower, typically *SULF*) and "full" yellow (yellow pigment spread across the face and upper petals, typical for *sulf* plants). The second axis stretches along the "red" manual score and describes the amount of the anthocyanin and its distribution governed mostly by two tightly linked loci *Rosea* and *Eluta*. The three red score subgroups can be characterised as "pale" (barely any magenta pigment, typical for *ros* plants), "patchy" (distinct, centrally located magenta patches on a paler background, typical for *ROSEL* plants) and "full" (homogeneous magenta coloration across the flower, with varying intensity described by the score, typical for *ROSel* plants). Although the parental types are "yellow" (full yellow, pale magenta) and "magenta" (pale yellow and full magenta), especially in hybrids the yellow and red scores can mix and match independently, forming white, pink, weak orange and dark orange (appearing red) phenotypes.

The machine image measurements from our pipeline (*image-stats*) described in chapter 4 on methods include a wide variety of variables and statistics on floral shape, size, colour and pattern. Most importantly, besides statistics for the whole flower our method distinguishes *subsets of pixels* classified based on presence and absence of pigments, using thresholds for hue and saturation. These subsets of pixels are *"magenta"* (only anthocyanin present) and *"widemagenta"* (some anthocyanin present), *"yellow"* (only aurone present) and *"wideyellow"* (some aurone present), *"orange"* (both anthocyanin and aurone present, the overlap of widemagenta and wideyellow) and *"white"* (very little pigment present), but also *"all"* (all pixels belonging to the flower) and *"other"* (not magenta, yellow, orange, nor white). For each of these colour categories we calculate the number of pixels in the corresponding subset and their proportion from all the flower pixels. In the *image-stats* table we include statistics summarising hue, saturation and value, as well as the light intensities of the classical three colour channels: red, green and blue over each of the pixel subsets. The statistics we include for each of these measurements are mean and variance, median, minimum, maximum, first and third quartile. If we used the classical definition of hue, where it ranges from red (0), through orange, yellow, green, cyan (0.5), blue, purple and magenta back towards red (1), in our flowers ranging from magenta through red and orange to yellow, we would face a problem. There would be a discontinuity around the red, which would correspond to 1 if approached from the magenta side, but would be 0 if approached from the orange side. This would mean, that with addition of just a little bit of aurone to the pure anthocyanin, the hue of the pixel would leap from 1 back down to 0, which would be very impractical for calculating and interpreting the hue statistics. Therefore, we decided to define a Hue (with capital H), such that $Hue = hue - round(hue)$ and include it in our measurements. This means, that the Hue stays the same on red-orange to yellow section (through green all the way to cyan ~ 0.5, but that is not really relevant) 0 to 0.5, and is decreased by 1 on the blue-purple through

magenta to red section, decreasing these values from $hue = 0.5 - 1$ to $Hue = -0.5 - 0$. This very simple transformation results in a measure of hue that is easily comparable to the classically defined hue *and* continuous on the hues of interest. Furthermore, we can talk about a shift of the colours "towards blue" or "towards yellow" based on decrease, or increase in Hue values, respectively.

Considering all statistics for all measures in all colour pixel subsets means that our dataset consists of more than 400 variables measured for all of the flowers. Therefore, in this chapter, we describe a meaningful, manual scores-informed dimension reduction put into the context of geographical location. For simple approximation of anthocyanin and aurone amounts in flowers we use *widemagenta index* and *wideyellow index*, first defined in chapter 4 on methods. These automatic colour indices are simply proportions of pigment containing pixels (widemagenta for all pixels containing anthocyanin and wideyellow for all aurone-containing pixels) multiplied by the mean pigment saturations in "clean" pixel subsets (magenta pixels for anthocyanin and yellow pixels for aurone).

We selected all individuals from the hybrid zone core which have both flower photographs and SNP genotypes available. To keep the camera setup consistent across the dataset, we selected all data collected in years 2016 - 2019, yielding $5,691$ individuals. We applied the image processing pipeline described in chapter 4 and ended up with four *image_stats* spreadsheets (one per season) stored in Barton group archive (folder *snapdragon/analyses/clines2021*) together with all data relevant to this chapter.

There are several kinds of errors. Besides those that happen outside in the field (a flower from a different, near-by plant is picked, or a wrong ID is typed into the system) we have very little power over, there are several errors that happen at the "photography desk" that can be, to some extent, controlled. Sometimes a mistake happens and a photograph is "incorrect": either the photograph does not depict the flower from a plant with the given ID, or the position of flower that it was supposed to. It can also happen that the colour of photographed flower does not match the manual scores due to human error. These mistakes can be either avoided altogether (only to some extent), fixed, or erroneous datapoints can be excluded from analysis, but all of these are quite tricky. Ideally, there would be a computer vision based "machine" checking for Plant IDs and flower positions (front, side, bottom, top) automatically. To check for errors in the manual scores, there is a "consistency script" that groups all flowers by their manual score that allows for colour consistency within the groups, single out and re-score the outliers.

Unfortunately, the computer vision machine does not yet exist and the consistency script has not been used consistently throughout the years. Therefore we are left with the last option: an attempt to exclude as many faulty datapoints from the analysis as possible. In order to do so, we aggregated data from all four years into one dataframe and then we grouped the plants by their manual red and yellow scores, compared the magenta and yellow indices within red and yellow score groups respectively and singled out the outliers (figure 5.1). After this procedure we were left with $5,542$ flower images we are using in all the later figures in this chapter, which is still large enough to result in substantial statistical power.

Since many of the flowers have no pixels belonging to some of the colour categories (in particular "white", "orange" and "other"), many of the machine colour measurement variables are not available (and marked as *NA* in the *image_stats*). In such cases, we set the saturation-related statistics (except for variance) for this colour pixel group to zero. However, we are still left with whole columns which are mostly *NA*s, which means they are quite useless in further statistical analysis. Therefore, we disregard the variables with more than 50 unavailable

measurements in further analysis, which leaves us with 112 machine colour measurement variables for all analysis in this chapter. The remaining measurements are mostly the statistics related to all and widemagenta pixel subsets and then number of pixels, pixel proportions and saturation-related statistics (except for variance) for all the other pixel groups. These include "yellow" and "wideyellow" pixel groups pointing to those areas of flowers where aurone is expressed.

5.1 Linear discriminant analysis

First, to better understand the complexity of the colour dataset, we decided to do Fisher's linear discriminant analysis (FLDA) [18] to find the linear combinations of machine measurements that capture the complex differences in flower coloration best.

Linear discriminant analysis (LDA) finds a direction defined by such a linear combination of variables in dataset, along which the datapoints separate best into two classes based on given labels. More specifically, it maximises the proportion of differences between group means to standard deviations within the groups. FLDA is a multidimensional generalisation of LDA with $n \geq 2$ groups, resulting into $n - 1$ such directions, i.e. linear discriminants, therefore we will sometimes refer to FLDA as LDA from now on. In this section, we will train the LDA to separate the differently-coloured flowers into their six manual phenotype categories using the multidimensional machine colour measurements. LDA considers variation *within the classes* and compares it to variation *between the classes*. This is in contrast to more widely known Principal components analysis (PCA), which seeks to maximise the overall variation in the dataset in general. However, parental phenotypes (yellow and magenta) are much more common than all of the hybrid phenotypes combined and therefore, PCA results would be subject to ascertainment bias.

The space described by the resulting five linear discriminants seems to separate the six phenotypes well (figure 5.2 and subfigures a) and b) from figure 5.4). The first discriminant (LD1) explaining 75.15% variation describes the amount of magenta pigment in flowers. LD3 then further separates the plants into pale vs. full yellow within their red score categories. However, what is LD2 and why is the cloud of datapoints V-shaped in the space defined by the first two linear discriminants? It *seems* to separate the intermediate manual red scores from both the low and the high ones. Since the intermediate red scores are more patchy than the low and high scores, LD2 could describe something like "patchiness", or, in other words: heterogeneity in magenta coloration across the flower face. But is it?

We can verify this hypothesis by studying different compositions of the individual LDs. All LDs are linear combinations of the machine colour measurement variables, with importance of each variable given by a scalar that differs for each of the machine colour measurement variables. If LD2 indeed describes patchiness, then on one hand, we would expect the scalings for colour *variance* be of larger magnitude in LD2 than in other LDs. On the other hand, and more importantly, within the LD2 scalings, we would expect the variance variables (as well distribution describing variables such as quartiles and medians) connected with the amount of magenta have the larger scalings than the other variables.

In figure 5.3 we plotted scalings of machine colour measurement variables with top 25% most variable scalings across the LDs. The scalings where the LD2 scaling is at either extreme of the scalings for the five LDs are: *variance of Hue in all pixels*, *variance of Hue in widemagenta pixels*, variance, median and third quartile of value in widemagenta pixels. However, in none of

these is the LD2 scaling of the highest magnitude, so LD2 might not be the only LD describing "patchiness". On the other hand, if we order the scalings of LD2, the widemagenta value and hue distribution-related variables are well represented in top 20 (there is a sharp drop in absolute value of scalings after 20).

To understand how the measurements interact with a human mind, we calculated correlations of some human-readable variables with the five LDs in table 5.2. Here, we can see that the LD2 is best correlated with variance of Hue in all pixels and in widemagenta pixels. This would be in favour of the patchiness hypothesis, as low Hue corresponds to shades of red closer to blue, i.e. magenta and higher Hue represents those closer to yellow, i.e. orange and gold. In plants with both anthocyanin and aurone present the Hue varies a lot between the patches with anthocyanin present and those low in anthocyanin.

	LD1	LD2	LD3	LD4	LD5
all_H_mean	0.97	-0.03	0.10	-0.04	-0.10
all_H_var	-0.13	0.66	-0.09	-0.18	-0.05
all_s_mean	-0.68	-0.44	0.43	0.11	-0.06
all_s_var	0.24	-0.16	0.05	0.34	0.11
magenta_proportion	-0.97	-0.04	-0.06	-0.01	0.13
yellow_proportion	0.91	-0.36	0.05	-0.03	0.07
widemagenta_proportion	-0.91	0.36	-0.05	0.03	-0.07
widemagenta_H_mean	0.95	0.00	0.13	-0.02	-0.09
widemagenta_H_var	-0.13	0.65	-0.02	-0.17	-0.02
widemagenta_s_mean	-0.69	-0.37	0.38	0.10	-0.06
widemagenta_s_var	0.28	-0.17	0.07	0.24	0.12
widemagenta_v_mean	0.38	0.26	-0.24	-0.01	0.03
widemagenta_v_var	-0.20	-0.22	0.05	0.05	-0.03
wideyellow_proportion	0.97	0.03	0.09	0.03	-0.13

Table 5.2: Linear discriminants separating the six phenotypic categories based on manual scores. In concordance with our observations in figure 5.2: the LD1 is dominated by measurements pointing to amounts of magenta pigment (all_H_mean also separates the flowers by their overall hue) and the LD2 separates flowers by magenta distribution pattern into magenta smooth (pale or full) and magenta heterogeneous being dominated by variance-related measures. The slight observed separation along the yellow axis by LD3 is not really visible.

5.1.1 The genetic basis of flower colour

To understand the phenotypic effects of the nine clinal SNPs we decided to find the differences between the two homozygous types of plants, i.e. plants with states 0 vs. 2 for each SNP (the 0-types are typical for *A. majus striatum* and the 2-types for *A. majus pseudomajus*). To do this, we did nine separate LDAs with the labels corresponding to the SNP value, always omitting the heterozygotes. The results are in table 5.3.

However, all of the nine SNPs are clinal and strongly linked to the two parental types. SNP2 ("third peak") and 3 (*Eluta*) are very tightly linked to SNP1 driving most of the variation in amount of anthocyanin, explaining, in turn, most of the variation in flower colour. Therefore, we decided to work with heterozygotes in SNP1 (and hence, most likely hybrids of the two parental types) for the other eight SNPs. Since there were not enough data for SNP5, we excluded it from the analysis.

	SNP1 *Ros1*	SNP2* *3rd peak*	SNP3* *El*	SNP4* *Sulf A*	SNP6* *Fla A*	SNP7* *Fla B*	SNP8* *Rubia*	SNP9* *Cremosa*
all_H_mean	-0.96	-0.14	-0.59	-0.63	-0.75	-0.39	-0.39	-0.59
all_H_var	0.19	-0.15	-0.08	0.24	0.24	0.27	0.27	0.19
all_s_mean	0.61	0.28	0.64	-0.26	-0.21	-0.09	-0.09	-0.22
all_s_var	-0.24	-0.06	-0.05	-0.22	-0.28	-0.12	-0.12	-0.21
magenta_proportion	0.96	0.13	0.61	0.60	0.73	0.36	0.36	0.58
yellow_proportion	-0.93	-0.17	-0.44	-0.45	-0.53	-0.39	-0.39	-0.44
widemagenta_proportion	0.93	0.17	0.45	0.45	0.53	0.39	0.39	0.43
widemagenta_H_mean	-0.93	-0.12	-0.57	-0.62	-0.76	-0.39	-0.39	-0.60
widemagenta_H_var	0.20	-0.14	-0.05	0.21	0.23	0.28	0.28	0.19
widemagenta_s_mean	0.63	0.26	0.63	-0.21	-0.16	-0.09	-0.09	-0.19
widemagenta_s_var	-0.29	-0.08	-0.10	-0.22	-0.27	-0.18	-0.18	-0.26
widemagenta_v_mean	-0.36	-0.24	-0.37	-0.04	-0.14	0.07	0.07	0.01
widemagenta_v_var	0.15	0.11	0.26	-0.14	-0.07	0.05	0.05	-0.04
wideyellow_proportion	-0.95	-0.12	-0.59	-0.61	-0.75	-0.36	-0.36	-0.60

Table 5.3: Correlations of discriminants derived from LDA separating homozygotes in each SNP separately. All SNPs are clinal and thus related to the two parental types and therefore, the LDA on SNPs denoted with an asterisk was performed using only $1,473$ plants heterozygous in SNP1.

As expected, the discriminant dividing ROS1 homozygotes (and hence, the two parental types by proxy) is heavily correlated with mean Hue of all and widemagenta pixels as well as magenta, widemagenta, yellow and wideyellow proportions.

We would expect *Eluta* SNP3 to be correlated with the variance variables, as the *striatum* form of *Eluta* locus (represented by 0) constricts magenta pigmentation and causes patchiness in flowers that contain anthocyanin. Since we constrained our analysis to SNP1 heterozygous plants only, all flowers should produce enough anthocyanin for *Eluta* to manifest itself, i.e. the *pseudomajus*-like 2-types should be homogeneously magenta pigmented and the *striatum*-like 0-types should be patchy. In plants that also contain aurone this means that the SNP3 is anticorrelated with Hue means in all and widemagenta pixels. The *pseudomajus* form of *Eluta* increases the proportion of magenta and widemagenta pixels. Interestingly, the correlation with variance-related variables is not as strong.

In *A. majus striatum Sulf* and *Fla* are associated with production of aurone. The *pseudomajus* form of aurone-related are therefore anticorrelated with mean Hue as well as with yellow and wideyellow pixel proportion.

As we already mentioned, this analysis suffers from linkage between the loci, but it can also be biased simply by the loci being present in one of the subspecies and not in the the other, which we attempted to solve by only including *Ros1* heterozygotes in the analysis. Another problem is that presumably the SNPs are not causal themselves and are only linked to the causal locus, which brings the correlation arbitrarily closer to zero.

5.2 The cline

To study phenotypic clines in the hybrid zone in this section we present both human-designed and data-derived one-dimensional continuous measures of the phenotype of $5,000$ plants

enriched in individuals growing in the hybrid zone core. Here we analyse the relationship of phenotypes to the W and Z system rather than the *easting* and *northing* system. The *easting* and *northing* system determines geographical position in metres. It is often used in works about the *Antirrhinum* hybrid zone in Planoles, as the hybrid zone does roughly stretch from west to east. However, more accurately it actually stretches along the roads, as the *Antirrhinum* plants are poor competitors and grow mostly on rocks and in disturbed habitats along the roads. Here, the W determines the position of the plants along the transect curve, which is a spline copying shape of the two main roads in the hybrid zone and the Z determines orthogonal distance from transect curve with the plus or minus sign giving the direction. Here we plot the clines relative to W, i.e. the main hybrid zone direction.

Since there are many more plants in the dataset coming from the centre of the hybrid zone than on the edges, it is difficult to see the patterns on individual plants in the core. Therefore we divided the dataset into 10 groups with approximately 560 plants each, based on increasing W (we first sorted the plants by W and then we assigned first tenth to the first W group, second tenth to the second W group, etc.).

To compare the colours of the plants with their W in the hybrid zone, we plotted the plants in the same colour LD space, but this time with colour of the marker signifying which W group the plant came from in figure 5.4. Comparing subfigures c) and d) with a) and b): focusing on the first, black to dark blue groups in the west side of the hybrid zone with low W, we can see that the markers are predominantly at the same place where the yellow plants used to be. Similarly, the dark red to red markers corresponding east-most, high W groups are concentrated in the area of the full magenta flowers. But what happened to the recombinant white, pink, orange and dark red phenotypes? If we check against the W group plots, we can see that these areas are now populated by cyan, green, yellow and golden markers, or the other way around: the highest concentration of cyan, green and yellow markers in positional plots coincide with those areas rich in recombinant phenotypes, which means that these phenotypes are indeed concentrated in the central parts of the hybrid zone.

The human-designed phenotypic measures we use are the widemagenta and wideyellow index based on the prior knowledge of the hybrid zone. It has been observed that the amount of magenta pigment visible in flowers increases and the amount of yellow pigment decreases as we move from west to east across the hybrid zone, switching from side rich in yellow *A. majus striatum* to magenta *A. majus pseudomajus* dominated area. In figure 5.5 we plotted boxplots for the magenta and yellow index within the ten W groups.

The clines are apparent in both colour indices and have about the same maximal steepness hinting at very similar cline width of about $400 - 500$ metres and centre around $W = 4700$ metres. However, while the magenta index starts increasing between the third and eight group, yellow seems to start decreasing slightly earlier: already between the second and eight W group. Interestingly, this difference between clines of *A. m. striatum* (yellow) and *A. m. pseudomajus* (magenta) haplotypes has already been observed [68].

There seem to be a lot of outliers concentrated at the outer sides of the hybrid zones. This is caused by the fact that the sides of the hybrid zone are mostly populated by the heavily uniform parental types filling the interquartile range (the "box") and heavily loading down the "whiskers", making every flower that diverges from the parental type an outlier.

5.2.1 The linear predictor

To take this one step further, we decided to describe the cline with a data-driven variable that would be a linear combination of machine colour measurement variables. To find out more about the relationship of geography and phenotype, we decided to define this quantity as the best linear predictor of W linearly combining the machine colour measurements. This can be obtained simply by doing a classical linear regression of the machine colour measurements on the W.

We plotted the boxplots for values of the resulting linear predictor of W in figure 5.6. As expected, the result is not as flat at the ends as the colour measurements, as this linear predictor of W is indeed doing a bit better job predicting W than the magenta and yellow indices above.

To compare the linear predictor of W with the six phenotype categories and the real W groups we plotted the flowers again in the same 3D linear discriminant space in figure 5.7.

Finally, to compare the three clines we plotted them all on the background of all the flowers coloured using the six phenotype categories based on manual scores in figure 5.8. All three seem to have similar width and centre.

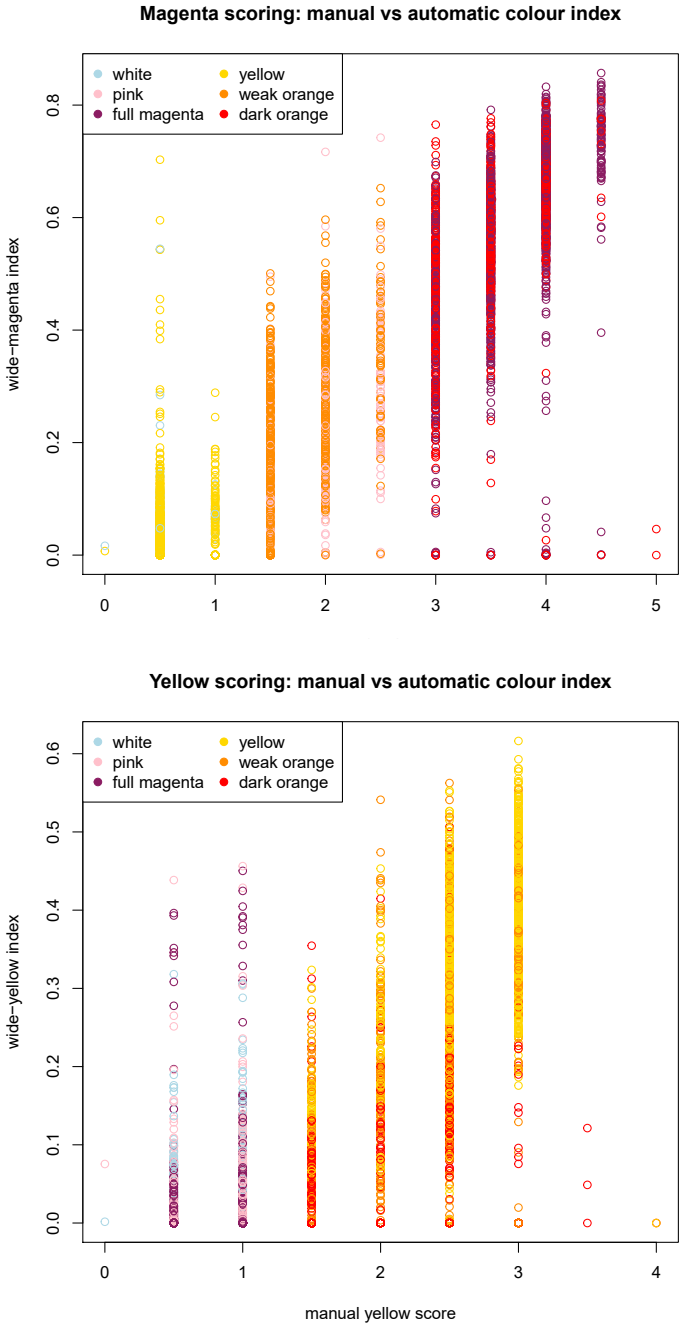


Figure 5.1: Comparison of manual scoring to the automatic colour indices with the six phenotype categories colour-coded. Although in general there is a strong correlation between automatic colour indices and manual scores, the presence of outliers is apparent especially for the very low and very high scores. From practice we can say, that high yellow score is sometimes overlooked in plants that are also high in anthocyanins, which explains the large number of flowers miscategorised as pale yellow "magenta" instead of "dark orange", although their wideyellow index is high.

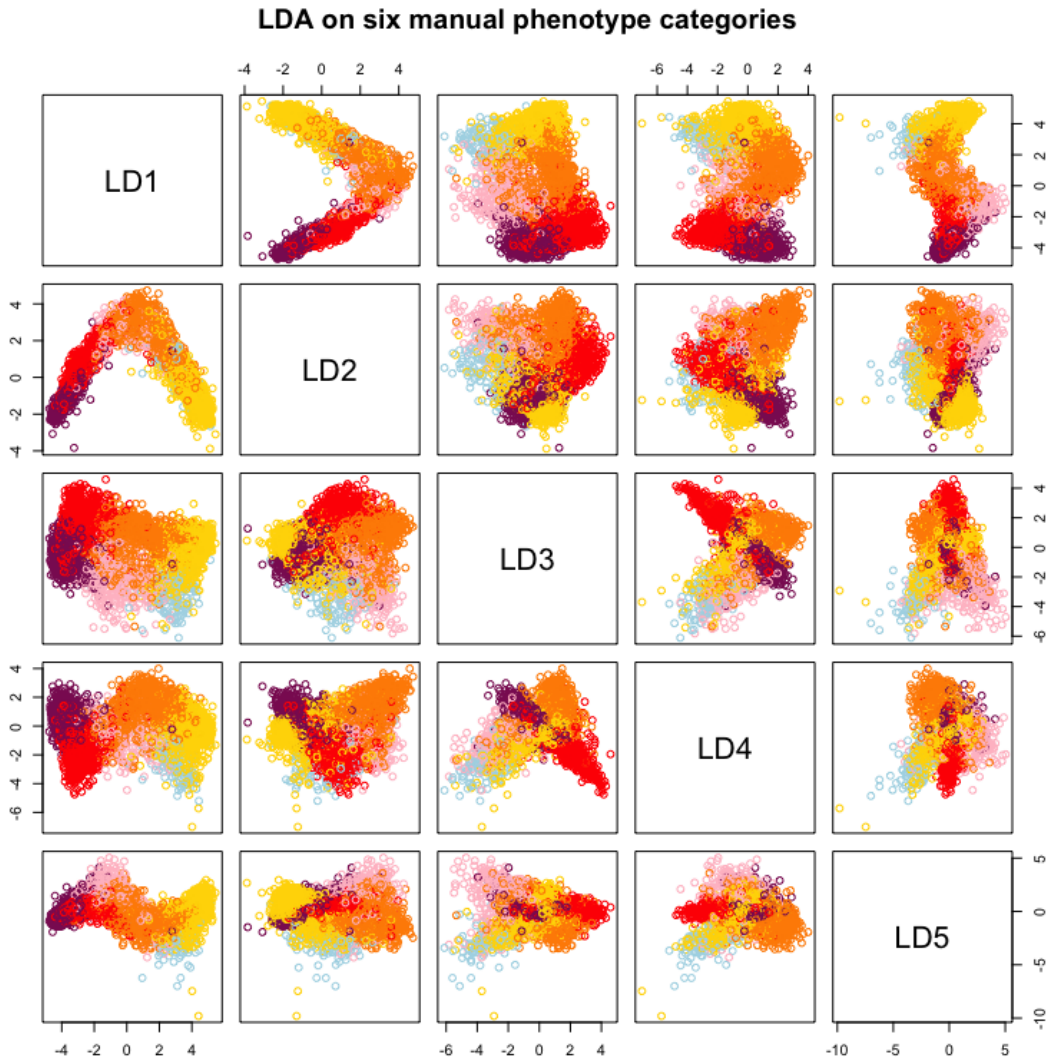


Figure 5.2: All components of Fisher linear discriminant analysis with the six classes corresponding to the six phenotype categories derived from manual scores. The proportions of explained variation are LD1: 75.15%, LD2: 11.16%, LD3: 6.86%, LD4: 4.37% and LD5: 2.46%.

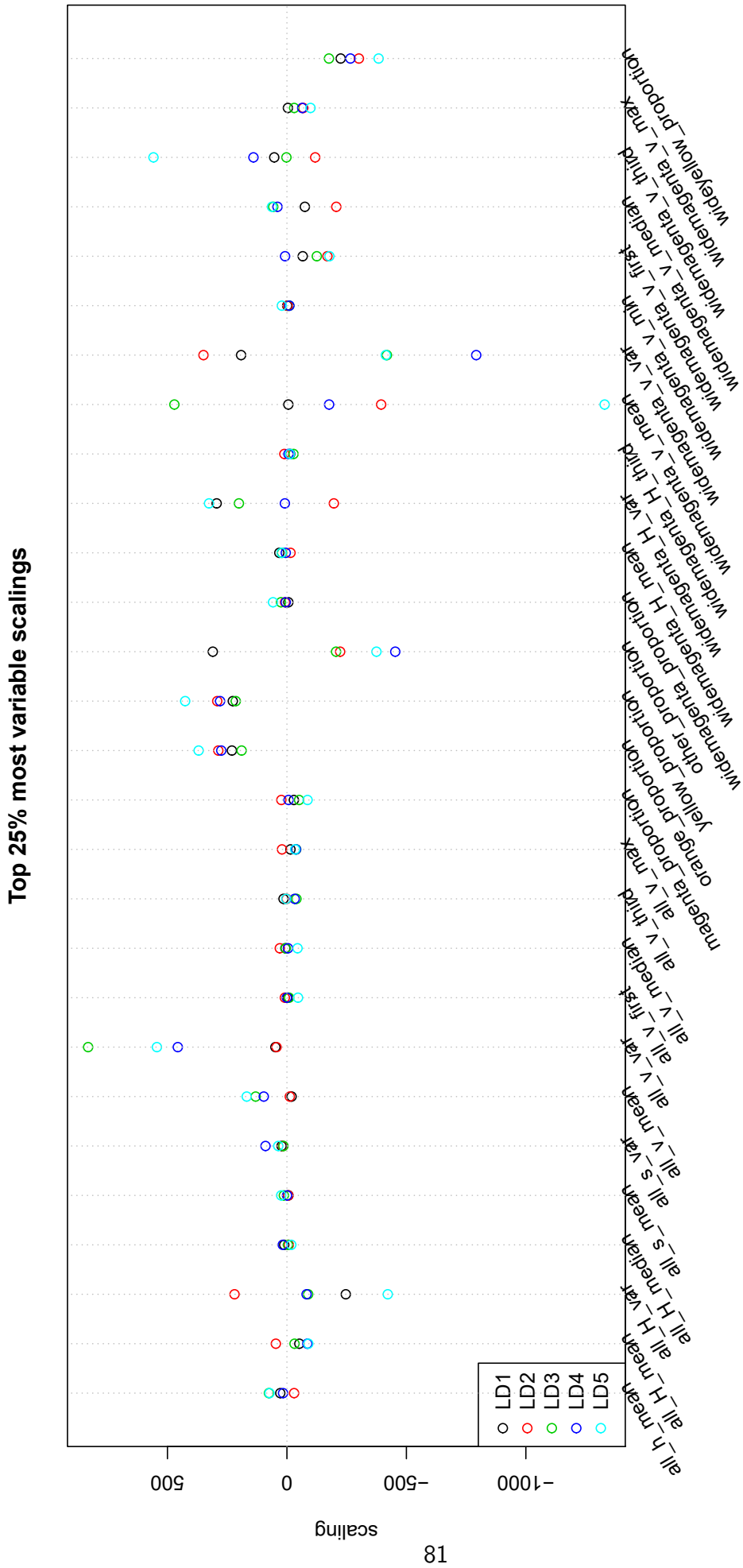


Figure 5.3: Comparison of the top 25% most variable scalings within the five LDs.

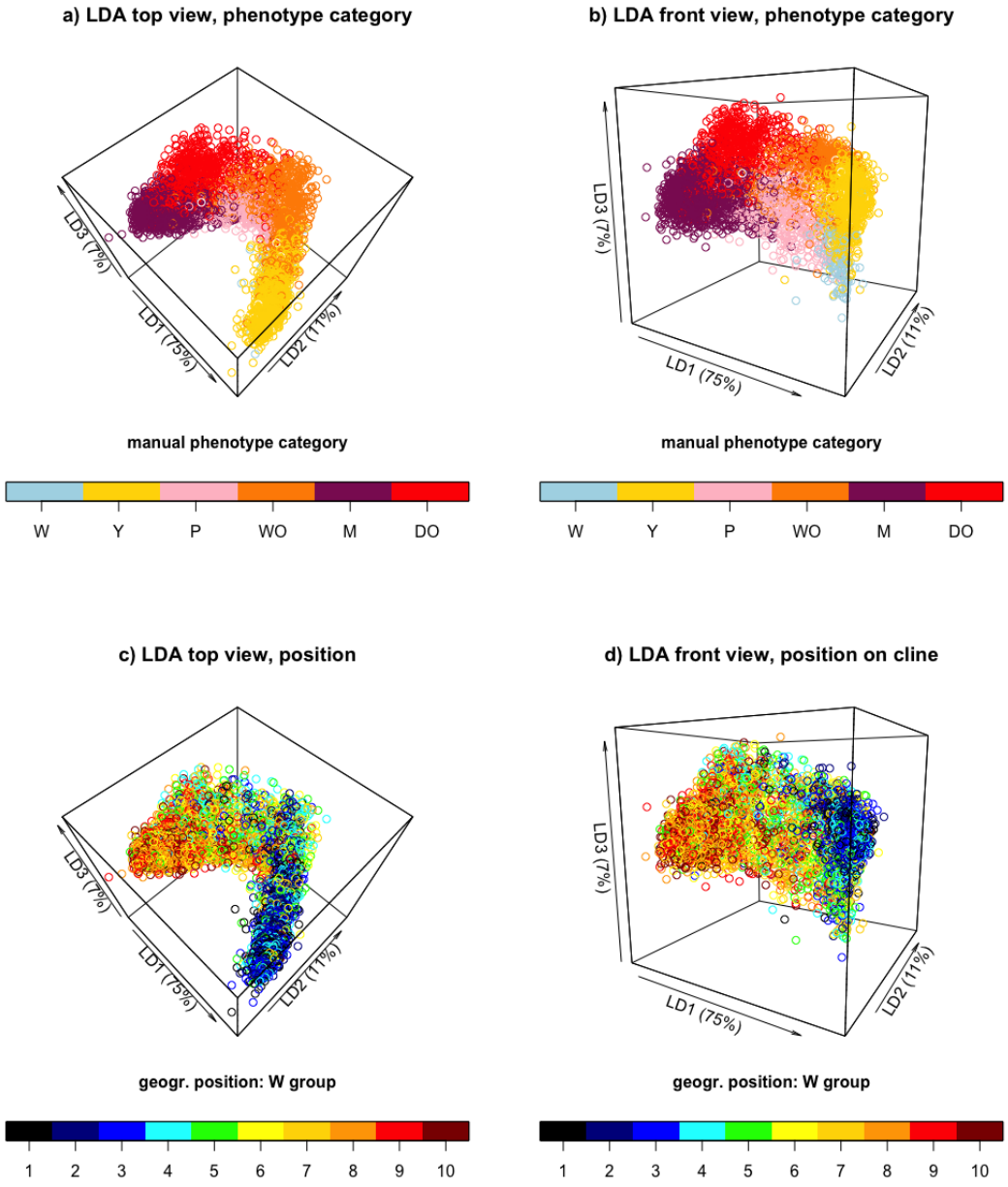


Figure 5.4: The 3D view of the automatic colour space in first three LDs. The markers represent plants coloured by manual phenotype category (a and b) and by their position across the cline and assignment to one of ten groups based on W (c and d). While the six phenotype category groups are clearly separated already in first three dimensions, the separation of plants based on their position on cline is much less apparent, although there is a clear trend.

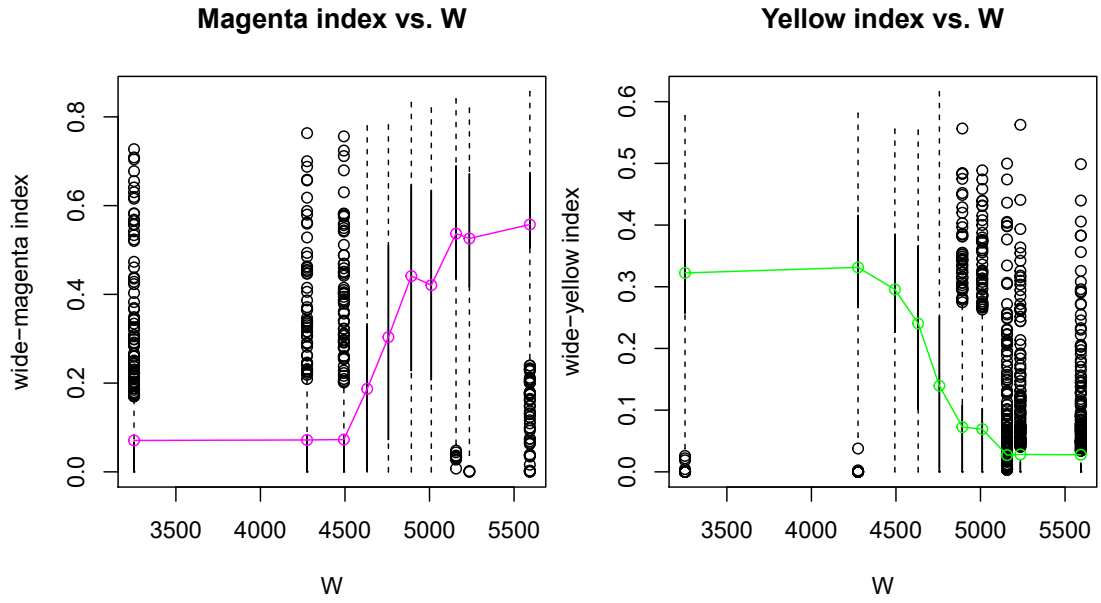


Figure 5.5: Boxplots for the magenta and yellow index within the ten W groups. The solid parts ("boxes") of the boxplots show the area between the first and the third quartile, the dashed whiskers extend to the last datapoint that is no further from the box than 1.5 times the interquartile range (i.e. the length of the box). All datapoints outside of the whiskers are considered outliers and plotted separately. In magenta and green coloured markers connected with lines are the magenta and yellow index means within the ten W groups.

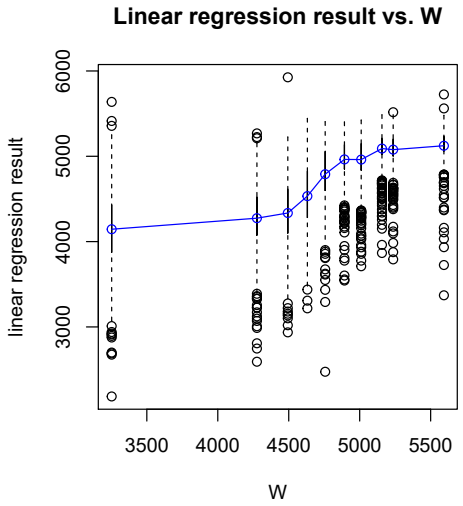


Figure 5.6: Boxplots for the linear predictor of W within the ten W groups. The blue coloured markers connected with a line show the linear predictor means within the ten W groups.

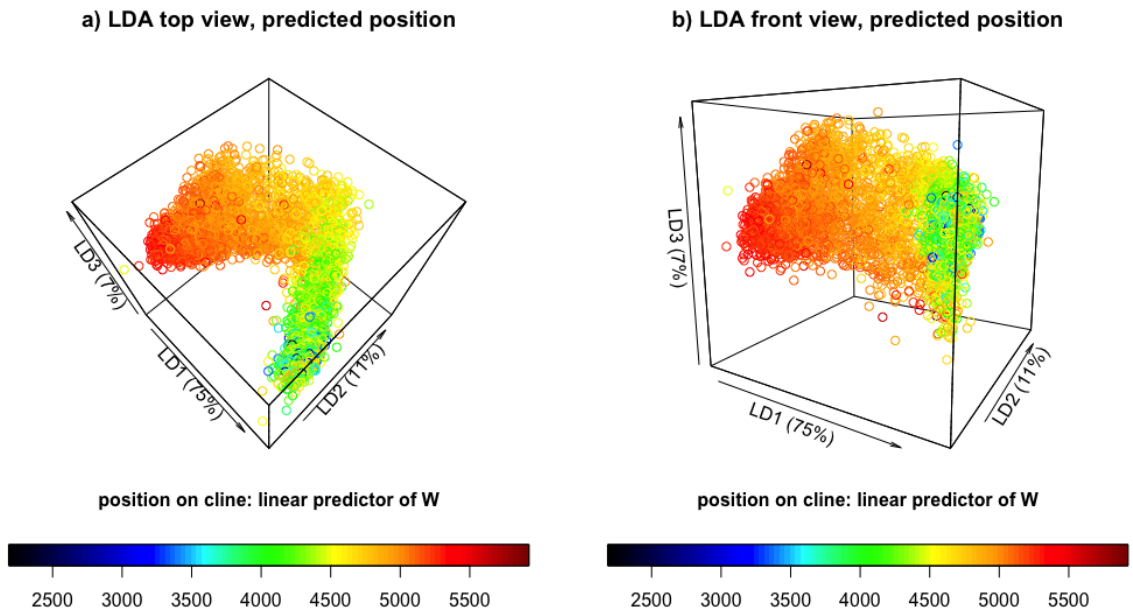


Figure 5.7: The 3D view of the automatic colour space in first three LDs, the markers representing plants are in the same positions as in figure 5.4, coloured by predicted W. Compared to colouring by the ten W groups (subfigures c) and d) in figure 5.4), colouring by continuous linear predictor of W in the 3D colour space is much smoother.

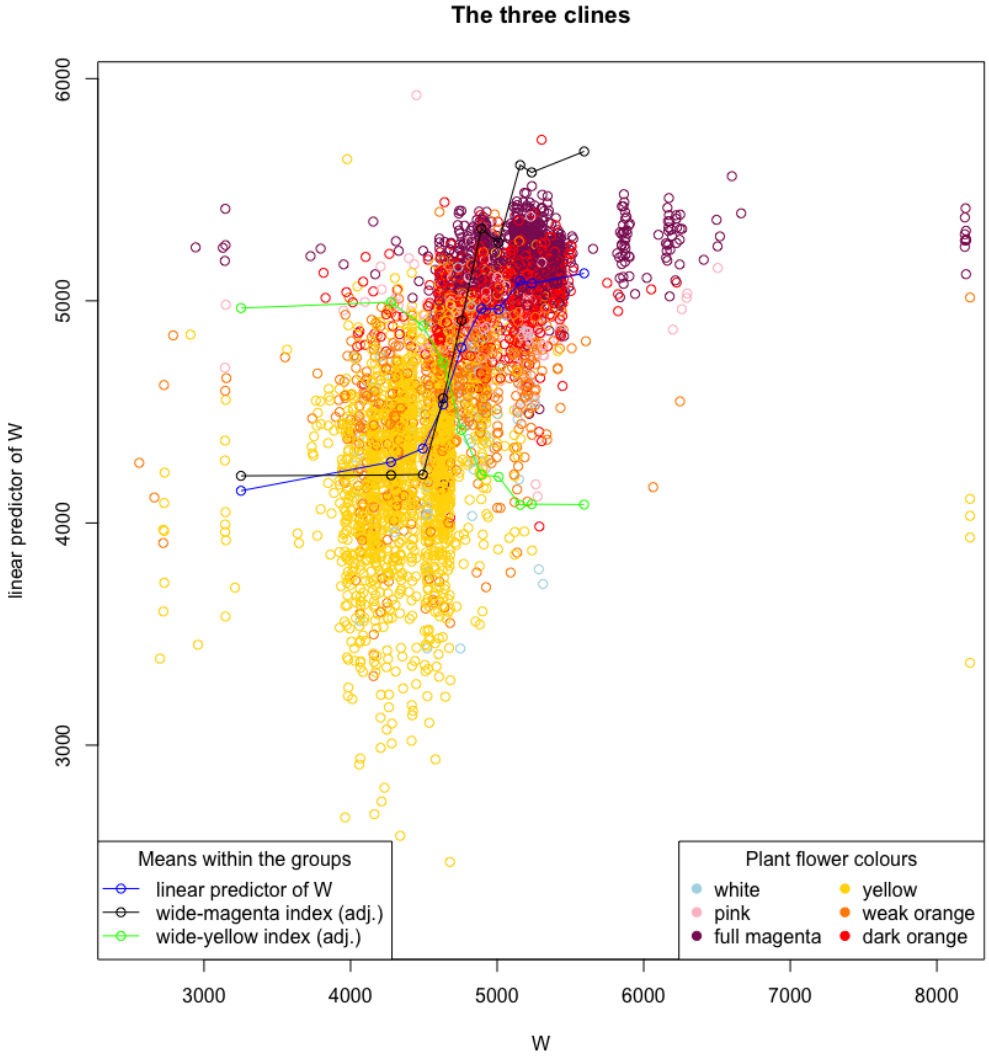


Figure 5.8: The three clines on background of all flowers in the dataset (the magenta and yellow index clines are linearly adjusted to fit in the image). The colours of markers correspond to phenotype category based on manual scores.

CHAPTER 6

Discussion

In this thesis, we developed methods suitable for large datasets of genomes and images, striving to account for their complex nature, while minimising human bias. We used these methods on a dataset of more than $20,000$ plant SNP genomes and corresponding flower images from a hybrid zone of two distinctly coloured subspecies of *Antirrhinum majus* to improve our understanding of the genetic nature of the flower colour in our study system.

6.1 Summary of results

Firstly, we used the advantage of large numbers of genotyped plants to estimate the partial haplotypes of the studied plants in chapter 2. Here, we focused on the region known to include two closely linked main loci (*Ros* and *El*) explaining the majority of variation in magenta coloration. We studied colour- and geography-related characteristics of the estimated haplotypes and how they connect to their relatedness; in chapter 3 we focused on the *Ros/El* recombinant types. We confirmed that there is a significant deficit of the double recessive recombinant (*rosel*) and quite surprisingly, we found that although present in larger numbers in the dataset, the haplotypes found in one of the parental types (magenta *A. m. pseudomajus*, *ROSel*) are genetically much less variable than those found in the other parental type (yellow *A. m. striatum*, *rosEL*).

In the second part of the thesis we focused on using automatic processing of flower images to extract meaningful characteristics from them. In chapter 4 we developed a pipeline capable of processing ten of thousands of images without human interaction and summarising each image into a handful of informative scores. In chapter 5 we went on to show the compatibility of these machine-calculated flower colour scores with the manual ones, and used them to study flower colour phenotypic clines in the hybrid zone of *A. majus*. Just as the phenotypic clines defined by manual flower colour scores are concordant with the genotypic ones (Parvathy Surendranadh, pers. comm.), we showed the same goes for the phenotypic clines defined by the machine colour scores.

6.2 Suggested extensions

Here, we suggest a couple of technical extensions that would help us to use the data more efficiently and address the topic of genotype, phenotype and fitness in the context of flower

colour more fully. Suggestions for further research stemming from findings in this thesis and curiosity of the author can be found in section 6.3.

6.2.1 A more robust pipeline

At the moment, the pipeline can deal with images from 2015 and later. The reason for this is that the images from earlier years vary in size, quality, format and camera that was used; this variation should be incorporated into the pipeline. Since digital cameras tend to compress the images and automatically colour-balance them in unpredictable ways, ideally, raw images should be used instead of JPEGs. Furthermore, in some years, the images are rotated randomly, which can be corrected for, either using metadata of the photograph, or by machine learning.

Another issue stopping us from using the full potential is human error, where the scorer writes a wrong photo code identifying the image into the sheet. This makes the photograph path identification part of the pipeline fail, as it does not find the photograph where indicated, rendering the photograph file invisible for downstream analysis. Similarly, the code can point to a photograph of a different object; either a different plant, or an image from a different angle. These mistakes are easy to make and tend to occur in groups. For example, not noting a change of date can lead to a faulty code in a full day of photographs and just skipping one code can propagate the chain of misaligned code-to-photograph events until the scorer notices.

Therefore, we suggest to automatically extract and check the plant ID, which is typically present and well visible in the photograph. Completing this relatively simple task could help us to not only use more of the photographs in the dataset, but also to avoid using the wrong images for the image analysis.

6.2.2 Pigment distribution

Artificial intelligence, also known as machine learning, comes in many forms and flavours. Similar to classical statistical model-based approaches such as, say, linear regression, it typically uses input data with desired labels (or values) for each data point to fine-tune the parameters in the algorithm such that the errors in assigning labels are optimised.

The two classical classes of machine learning algorithms would be *supervised*, where the "ground truth" labels are indeed supplied in the input data and *unsupervised*, where the labels are not available and the algorithm has to rely on intrinsic characteristics of the data and structure in the dataset, for example by using clustering.

In the process of fine-tuning the parameters (i.e. *training* the algorithm), the input dataset is often split into a *training* and *testing* dataset to make sure that the algorithm is not biased, but it is universal. In other words, to make sure that the algorithm is not *over-fitted* to one concrete training dataset.

As opposed to classical statistical methods which are concerned with "goodness of fit" of the model to the data and interpretability of the parameters as well as accurate prediction, the machine learning models tend to be too complex to understand the role of individual parameters and the focus lies heavily on optimising the prediction. Although this may seem a shortcoming of machine learning algorithms, it also means that they can deal with much more complex and convoluted problems or in situations, where the conditions for classical statistical approaches simply cannot be fulfilled. Therefore, machine learning approaches are particularly useful when it comes to image processing.

In chapter 4, we used machine learning to identify the position of the flower in the photograph. To train the algorithm we used an input dataset consisting of several photographs with flowers of different colours with paired "labels", i.e. matrices of the same size of the input image, with ones where the flowers are and zeroes everywhere else. This process is followed by colour quantification of the flower pixels and although the image analysis pipeline outputs more than 30 colour characteristics for each image, these are still selected by a human and there may be biologically meaningful variation beyond these measurements.

In particular, the pipeline does not yet provide much information on distribution of the pigment apart from variance of colour characteristics across the flower pixels. Thus, we are losing many *dimensions* of information about the floral appearance. However, the distribution of pigment and possible patterns it forms may be particularly important, as pollinators are capable of picking up differences in pigment distribution. Moreover, the colour patterns can create pollination guides, such as a yellow patch at the flower entrance in *A. m. pseudomajus*, or internal magenta venation in *A. m. striatum* assisting the pollinators to find their way into the flower [7]. Thus, the colour pattern could have an impact on pollination success and, ultimately, on the fitness of the plant. Therefore, we believe the pigment distribution should be studied and the pipeline should be extended to allow it.

One way to go about it would be to use machine learning to define the *pixels belonging to respective separate flower parts* for each flower, as suggested in figure 4.15 in chapter 4 and then to define the measurements for each flower part separately. In other words, we would like to end up with four mask matrices for each flower image, one mask matrix indicating the pixels belonging to the upper petals, another mask matrix showing where the pixels belonging to the lower petals are, etc. This would be a conceptually simple approach with well interpretable results. Another advantage is that the tools already exist (such as Mask R-CNN, [25]), are implemented in Python and well established in the computer vision community. The shortcoming may be that this approach would require a training dataset in the order of tens, maybe even hundreds of annotated flower images (consisting of four matrices per image as described above). However, in comparison to the annual amount of field work, this would be a relatively inexpensive one-time investment.

Nonetheless, even though the resulting flower colour characteristics would be biologically informed, they would still be human-defined and thus may still lack some of the recurring patterns that are not obvious to the human mind. Therefore, one could try to let the artificial intelligence to pick up on the flower traits by itself (i.e. *unsupervised* learning), finding clusters of flowers with similar traits in the sea of multidimensional images. On one hand, this would remove the human bias, but on the other hand, the results may not be interpretable and the amount of work invested in such a project would not be guaranteed a reward in the form of new and clear findings.

Alternatively, one could try to train a machine learning algorithm to predict the most likely appearance of a flower from the plant genotype and then reverse-engineer the predictions, which could lead to discovery of new loci with impact on floral traits. Or vice-versa, one could train an algorithm to predict most likely genotype from an image, which could help to approximately "genotype" the plant quickly and inexpensively in the field, just using the photograph.

6.2.3 Pollinator perception

So far, the image analysis was performed such that it fit the nature of the human eye-sight and the capabilities of a standard digital camera. However, it has been shown that bumblebees, who are the main pollinators in our system, have three types of receptors with sensitivity peaks in UV, blue and green part of the spectrum [56] as opposed to the trichromatic red, green and blue combination typical for the human eye. Luckily, the method to transform reflectance light spectra into a perception vision of various pollinators including bumblebee is available and by now well established [59]. Therefore we suggest transforming the measures of visible light into the bumblebee system whenever any conclusions about pollinator behaviour are to take place.

While this should be reasonably easy to do for blue and green parts of the spectrum, unfortunately, the UV reflectance of flowers is not captured by a standard camera and thus the data on UV is not available for the existent dataset used here (although it may be relevant).

Another relevant feature possibly affecting bumblebee behaviour not captured in the image dataset is the presence of volatile substances in the flowers, particularly, as bumblebees have a very well developed sense of smell and use it for foraging decisions. It is quite possible that the scent-based decisions contribute to the reproductive isolation, as is the case in hybrid zone of two *Mimulus* species [10]. Coincidentally, there seems to be a difference in scent between the *A. m. pseudomajus* and *A. m. striatum* flowers obvious even to imperfect olfactory organs of a human photographer coming to contact with hundreds of flowers with different phenotypes daily. Unfortunately, it is out of scope of the available data to study this interesting phenomenon in more depth, but one should still be aware of it when drawing conclusions about pollinator behaviour.

6.2.4 More detailed haplotypes

At the moment, we are working mostly with genotypes on 120 SNPs (and some hundreds of genotypes on 240 SNPs). So far, we have estimated the haplotypes in the *Ros/El* locus found to be governing most of the floral anthocyanin production and distribution in our system, covered by 12 (or 24) SNPs and it seems to match the expected anthocyanin phenotypes quite well. The most notable differences from what was expected is perhaps that dominant *Eluta* allele only restricts magenta pigmentation in plants heterozygous in the *Rosea* locus. Moreover, it seems that there is a handful of partial haplotypes (limited to *Ros/El* region) that consistently lead to different phenotypes than expected (figure 3.3).

To understand the genetic basis of flower colour more fully, including possible other interesting loci, it would be interesting to see the *whole phased SNP genotypes*. Unfortunately, the expectation-maximisation algorithm in the form used here would not scale well with a number of SNPs much larger than the number of SNPs in the *Ros/El* region. As discussed in chapter 2, other computational methods do not deal well with the large variety of haplotypes that would be expected and downsampling the dataset just to fit does not seem ideal.

A technological solution to this problem seems to lie in sequencing and everyone who would argue that whole genome sequencing would offer a much higher level of detail than a few hundreds of SNPs per plant would be right. In fact, a haplotagging [43] dataset providing whole genome sequences, accurately phased (i.e. whole-genome sequence haplotypes) for several hundreds of *Antirrhinum* plants, is in preparation (Sean Stankowski, pers. comm.). Using whole-genome haplotypes can help us to understand the genetic variation, as well the mechanisms in which the identified loci affect the flower colour in more depth, especially if the

plants are carefully selected. More specifically, it could help us to understand how different are the outlier haplotypes from figure 3.3. However, the amount of storage and resources it would take would make it impractical to sequence the whole genome of every single plant in the hybrid zone. Therefore, while we acknowledge the advantages of sequencing, we insist that consistent, simple and inexpensive SNP genotyping has an important role in studying large wild populations.

For example, such SNP genotypes have been used in [15] to infer the subsets of seeds in same seed pod (coming from the same mother) that were pollinated by the same father in an experimental *A. majus* derived from the wild population studied here. In another example, the same dataset of genotypes as used in this thesis is used to derive a pedigree identifying parents-offspring trios in our dataset, further leading to estimates of fitness of the six basic flower colour phenotypes (David Field, pers. comm.).

6.3 Open questions

With the image analysis pipeline up and running and the data on images available, several questions are still open. Perhaps the most burning one is, whether we can reject the simplistic *null model* from section 1.4.2, where two loci with two alleles each completely explain the amount and distribution of anthocyanin. From our findings on recombinants in chapter 3 it is apparent that this null model is not sufficient to explain the variation in magenta pigmentation of flowers in our dataset, as *Eluta* only seems to affect the plants heterozygous at *Rosea* (figure 3.1) and there are several haplotypes consistently resulting in different phenotypes than expected (figure 3.3). However, more work is needed to quantify the variation that is not captured by the known loci, to define more loci and more alleles affecting the flower colour. Perhaps, this can be achieved by studying the *Ros/El* locus in more detail using the haplotagging dataset, or via genome-wide association studies.

On a similar note: we have shown that variation of floral pigments seems to be rather continuous, possibly affected by several loci and other factors and hence can be understood as a quantitative trait. This can lead us to questions like how heritable is flower colour, and how significant are the effects of non-heritable factors? These can be addressed for example by combining the image analysis data with the pedigree in parent-offspring regression, or better, the animal model [75]. Thus, one can directly estimate the components of variance in flower colour.

It is known that there are more agents affecting the flower pigmentation. Most notably, the amount of UV radiation has been shown to increase the production of UV-absorbing pigments in flowers [36]. Quite fortunately, the group managed to get access to accurate historical meteorological data in Planoles in summer 2021. Thus, the available image analysis can be combined with the newly acquired data on amount of daily sunshine and precipitation in Planoles to find an estimate of the effect of meteorological factors on flower colour.

Last but not least, the analysis in chapters 2 and 3 showed that the haplotypes in *Ros/El* region regulating most of the anthocyanin expression in the flower is much less variable in anthocyanin rich, full magenta *A. m. pseudomajus* plants than in yellow *A. m. striatum* plants. Does it mean that there is "just one way to be red" and the genetic region is under a strong purifying selection in the magenta *A. m. pseudomajus* population, while less restricted in the yellow parental type? Or, since we know that the *A. m. pseudomajus* population is expanding into the yellow population, does it mean that the *A. m. pseudomajus* haplotypes

are just much more closely related? This is indeed a very interesting question that could be answered, for example by comparing variability of the two parental populations in other genetic regions, especially those unrelated to flower colour. If the decrease in haplotype variation in the *Ros/El* region in *A. m. pseudomajus* plants is due to purifying selection, we would expect the decrease in haplotype variation in the same plants to be limited only to the *Ros/El* region and nowhere else. If, on the contrary, the haplotype variation was decreased across the whole genomes of *A. m. pseudomajus* plants compared to the other plants, we would conclude that the observed pattern is due to higher relatedness of the *A. m. pseudomajus* population near the hybrid zone, possibly due to its recent expansion. Another strategy to address this question would be to study haplotype variation in genetic regions controlling the expression of the other important pigment: the yellow aurone. Especially, one should check whether a similar pattern of decreased variation is present around the aurone-controlling loci in yellow plants. However, this might get complicated, as the main aurone-controlling locus *Sulfurea* is in fact a deletion of a small silencing RNA for which the *A. m. striatum* plants are homozygous. Thus, the *Sulfurea* sequence is not present at all in *A. m. striatum* plants and its haplotype variation in these plants cannot be studied directly.

6.4 Studying complex variation

Although we often see studies focusing on simple or even discrete traits in studies of colouration, the variation of visual phenotypes found in nature is often more complex, frequently continuous and high-dimensional. For example, the wing pattern of *Heliconius* butterflies is often summarised by a few Mendelian genotypes, although it is more complex.

Studying *continuous* traits can hardly be called a novelty. For example, the heredity of human height [19, 17] and milk yields of dairy cows [76] (among others) have been already well studied more than a hundred years ago, setting the foundations of Quantitative Genetics.

However, these traits, similar to many continuous traits studied nowadays, are one-dimensional and thus, in a sense of dimensionality, simple. No one is going to argue that height (or milk yield) are unimportant traits by themselves. However, the human height (maybe too) easily isolated from body shape, bone shape, length and mass tells us little about an individual's fitness in comparison to high-dimensional system it is a natural part of and the same goes for milk yield (quantity) compared to rather multifaceted milk quality.

Similarly, the red patch on a lepidopteran wing may or may not help to deter predators or to attract mates using their vision, but it certainly is an intrinsic part of a larger, visually perceived system: the overall appearance of the wing, i.e. presence of other coloured patches, size of the red patch, its intensity, etc. And finally, the same goes for flowers: in the world where the size, shape, pigmentation and its distribution pattern, odour and even electric field in flowers [11] all affect their chances of being pollinated (and where the pollen comes from) it seems unavoidable to move towards more complex and *complete* phenotypes.

Fortunately, the framework for quantitative understanding with multidimensional continuous traits has also been established in the form of Multivariate Quantitative Genetics. Selection on correlated multivariate traits is a famous problem formulated already by Darwin and later dealt with throughout the 20th century [37, 38] using the methods of multivariate statistics and following findings of Pearson more than hundred years ago [48].

An example of a field connected to Multivariate Quantitative Genetics [52] that has been historically striving to grasp real, complex and high dimensional phenotypes is Morphometrics,

studying shapes of organisms. Starting with correlating individual measurements, the entire field has moved on to more sophisticated, partially or fully automated technologies and methods and branched out to new, computational fields such as Computational Anatomy, exploiting machine learning and developing suitable statistical methods on the way [42].

Apart from statistical methods, we also need automated phenotyping to fully embrace the phenotypes in their complexity. Automated phenotyping is becoming a trend for plants, animal and for human samples, in academia and in industry (for example large-scale automated phenotyping of farmed plants with PlantEye by Phenospex [69]) alike. With large datasets available, complex questions in mind and large computational resources in hand, we decided to develop high-throughput automated pipeline especially suited for detecting fine differences in colouration of flowers in this thesis. To underline how timely our efforts are, there exist several other tools for automated measurements of plant colours from digital images published around the time this thesis was in preparation (for more detail see 4). However, unlike the pipeline developed here, all methods found elsewhere seem to use simple thresholding for separating the object of interest from background (except [20], that does not mention segmentation at all), which proved ineffective in the case of highly variable *A. majus* flowers coming from the hybrid zone.

6.4.1 Studying complex genotypes

Similar to phenotypes, the studied variation in genotypes is often reduced to a couple of isolated SNPs. However, in GWAS studies of complex traits, the most significant SNPs combined often explain only a small fraction of the predicted genetic variance [41], while the rest can be attributed to large numbers of SNPs with smaller effects, or to rare variants [78]. Furthermore, [6] argues, that rather than inside the protein-coding genes, large proportion of predicted genetic variance can be explained by non-coding regions.

Although many of these findings were published in the context of complex human disease, or notoriously polygenic traits such as human height or BMI, this is especially relevant, as wild populations are a source of highly variable genotypes and the observed flower colour phenotypes are also continuous and complex. Furthermore, as we have shown in figure 3.3 for haplotypes on SNPs in the anthocyanin-regulating *Ros/El* region, although the majority of phenotype variation can be explained by one or two SNPs, the different haplotypes may have a significant effect on the phenotype. Therefore, we argue that when trying to describe the whole picture, it may be beneficial to consider genotypic effects jointly and genotypes in as much detail and entirety as possible, especially when studying a wild population.

6.5 Closing remarks

An amazing amount of work has been done on the simple. Now, with the understanding, resources and methods at hand it is time to embrace the real, to see the phenotype and genotype variation in its entirety and let it speak for itself in a language we are beginning to understand. We hope that the efforts to do so in this thesis will be followed by many more.

Bibliography

[1] R. J. Abbott. Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *Journal of Systematics and Evolution*, 55(4):238–258, 2017.

[2] G. Aldridge and D. R. Campbell. Variation in pollinator preference between two ipomopsis contact sites that differ in hybridization rate. *Evolution*, 61(1):99–110, 2007.

[3] L. Arathoon, P. Surendranadh, N. Barton, D. L. Field, M. Pickup, and C. A. Baskett. Effects of fine-scale population structure on inbreeding in a long-term study of snapdragons (Antirrhinum majus). *bioRxiv*, pages 2020–08, 2021.

[4] N. H. Barton and G. M. Hewitt. Analysis of hybrid zones. *Annual review of Ecology and Systematics*, 16(1):113–148, 1985.

[5] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk. ilastik: interactive machine learning for (bio)image analysis. *Nature Methods*, Sept. 2019.

[6] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[7] D. Bradley, P. Xu, I.-I. Mohorianu, A. Whibley, D. Field, H. Tavares, M. Couchman, L. Copsey, R. Carpenter, M. Li, et al. Evolution of flower color pattern through selection on regulatory small RNAs. *Science*, 358(6365):925–928, 2017.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[10] K. J. Byers, H. Bradshaw Jr, and J. A. Riffell. Three floral volatiles contribute to differential pollinator attraction in monkeyflowers (Mimulus). *Journal of Experimental Biology*, 217(4):614–623, 2014.

[11] D. Clarke, H. Whitney, G. Sutton, and D. Robert. Detection and learning of floral electric fields by bumblebees. *Science*, 340(6128):66–69, 2013.

[12] E. S. Coen and E. M. Meyerowitz. The war of the whorls: genetic interactions controlling flower development. *Nature*, 353(6339):31, 1991.

[13] A. Davison, H. J. Jackson, E. W. Murphy, and T. Reader. Discrete or indiscrete? Redefining the colour polymorphism of the land snail Cepaea nemoralis. *Heredity*, 123(2):162–175, 2019.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[15] T. J. Ellis. *The role of pollinator-mediated selection in the maintenance of a flower colour polymorphism in an Antirrhinum majus hybrid zone*. PhD thesis, Institute of Science and Technology, Austria, 2016.

[16] J. Endler. Geographic variation, speciation, and clines princeton university press. *Princeton, New Jersey, USA*, 1977.

[17] R. A. Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[18] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[19] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[20] J. E. Garcia, A. D. Greentree, M. Shrestha, A. Dorin, and A. G. Dyer. Flower colours through the lens: quantitative measurement with visible and ultraviolet digital photography. *PloS one*, 9(5):e96646, 2014.

[21] V. Grant. Pollination systems as isolating mechanisms in angiosperms. *Evolution*, pages 82–97, 1949.

[22] J. Hackbarth, P. Michaelis, and G. Scheller. Untersuchungen an dem Antirrhinum-Wildsippen-Sortiment von E. Baur. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, 80(1):1–102, 1942.

[23] J. Haldane. The theory of a cline. *Journal of genetics*, 48(3):277–284, 1948.

[24] C. Handelman and J. R. Kohn. Hummingbird color preference within a natural hybrid population of Mimulus aurantiacus (Phrymaceae). *Plant Species Biology*, 29(1):65–72, 2014.

[25] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[27] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[28] A. Horridge. Bee vision of pattern and 3d. the bidder lecture 1994. *Bioessays*, 16(12):877–884, 1994.

[29] E. Huss, K. Bar Yosef, and M. Zaccai. Humans' relationship to flowers as an example of the multiple components of embodied aesthetics. *Behavioral Sciences*, 8(3):32, 2018.

[30] S. Istrail. Computing haplotype frequencies and haplotype phasing via the Expectation Maximization (EM) algorithm, October 2012.

[31] C. C. Jaworski, C. Andalo, C. Raynaud, V. Simon, C. Thébaud, and J. Chave. The influence of prior learning experience on pollinator choice: an experiment using bumblebees on two wild floral types of antirrhinum majus. *PloS one*, 10(8):e0130225, 2015.

[32] C. C. Jaworski, C. Thebaud, and J. Chave. Dynamics and persistence in a metacommunity centred on the plant Antirrhinum majus: theoretical predictions and an empirical test. *Journal of Ecology*, 104(2):456–468, 2016.

[33] C. D. Jiggins. *The ecology and evolution of Heliconius butterflies*. Oxford University Press, 2017.

[34] E. C. Jorgensen and T. Geissman. The chemistry of flower pigmentation in Antirrhinum majus color genotypes. iii. relative anthocyanin and aurone concentrations. *Archives of Biochemistry and Biophysics*, 55(2):389–402, 1955.

[35] D. Kendal, C. E. Hauser, G. E. Garrard, S. Jellinek, K. M. Giljohann, and J. L. Moore. Quantifying plant colour and colour difference as perceived by humans using digital images. *PLoS one*, 8(8):e72296, 2013.

[36] M. H. Koski, D. MacQueen, and T.-L. Ashman. Floral pigmentation has responded rapidly to global change in ozone and temperature. *Current Biology*, 30(22):4425–4431, 2020.

[37] R. Lande. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution*, pages 402–416, 1979.

[38] R. Lande and S. J. Arnold. The measurement of selection on correlated characters. *Evolution*, pages 1210–1226, 1983.

[39] M. Li, M. H. Frank, and Z. Migicovsky. Colourquant: a high-throughput technique to extract and quantify colour phenotypes from plant images. *arXiv preprint arXiv:1903.01652*, 2019.

[40] M. Li, D. Zhang, Q. Gao, Y. Luo, H. Zhang, B. Ma, C. Chen, A. Whibley, Y. Zhang, Y. Cao, et al. Genome structure and evolution of Antirrhinum majus. *Nature Plants*, 5(2):174–183, 2019.

[41] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[42] L. F. Marcus, M. Corti, A. Loy, G. J. Naylor, and D. E. Slice. *Advances in morphometrics*, volume 284. Springer Science & Business Media, 2013.

[43] J. I. Meier, P. A. Salazar, M. Kučka, R. W. Davies, A. Dréau, I. Aldás, O. B. Power, N. J. Nadeau, J. R. Bridle, C. Rolian, et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences*, 118(25), 2021.

[44] G. Mendel. Versuche über plflanzenhybriden. verhandlungen des naturforschenden vereines in brünn, bd. iv für das jahr 1865. *Abhandlungen*, pages 3–47, 1866.

[45] K. Niovi Jones and J. S. Reithel. Pollinator-mediated selection on a flower color polymorphism in experimental populations of Antirrhinum (Scrophulariaceae). *American Journal of Botany*, 88(3):447–454, 2001.

[46] B. Nürnberger, N. Barton, C. MacCallum, J. Gilchrist, and M. Appleby. Natural selection on quantitative traits in the Bombina hybrid zone. *Evolution*, 49(6):1224–1238, 1995.

[47] A. Orteu and C. D. Jiggins. The genomics of coloration provides insights into adaptive evolution. *Nature Reviews Genetics*, 21(8):461–475, 2020.

[48] K. Pearson. I. Mathematical contributions to the theory of evolution.—XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 200(321-330):1–66, 1903.

[49] K. J. Rankin, C. A. McLean, D. J. Kemp, and D. Stuart-Fox. The genetic basis of discrete and quantitative colour variation in the polymorphic lizard, Ctenophorus decresii. *BMC Evolutionary Biology*, 16(1):1–14, 2016.

[50] H. Ringbauer. Antirrhinum majus hybridzone in the Ribes Valley: Historical records. *A literature review*, 2016.

[51] H. Ringbauer, A. Kolesnikov, D. Field, and N. Barton. Estimating barriers to gene flow from distorted isolation-by-distance patterns. *Genetics*, 208(3):1231–1245, 2018.

[52] B. Riska. Some models for development, growth, and morphometric correlation. *Evolution*, 40(6):1303–1311, 1986.

[53] C. Rother, V. Kolmogorov, and A. Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[54] K. Schwinn, J. Venail, Y. Shang, S. Mackay, V. Alm, E. Butelli, R. Oyama, P. Bailey, K. Davies, and C. Martin. A small family of myb-regulatory genes controls floral pigmentation intensity and patterning in the genus antirrhinum. *The Plant Cell*, 18(4):831–851, 2006.

[55] Y. Shang, J. Venail, S. Mackay, P. C. Bailey, K. E. Schwinn, P. E. Jameson, C. R. Martin, and K. M. Davies. The molecular basis for venation patterning of pigmentation and its effect on pollinator attraction in flowers of Antirrhinum. *New Phytologist*, 189(2):602–615, 2011.

[56] P. Skorupski, T. F. Döring, and L. Chittka. Photoreceptor spectral sensitivity in island and mainland populations of the bumblebee, Bombus terrestris. *Journal of Comparative Physiology A*, 193(5):485–494, 2007.

[57] M. Slatkin. Gene flow and selection in a cline. *Genetics*, 75(4):733–756, 1973.

[58] J. M. Sobel and M. A. Streisfeld. Flower color as a model system for studies of plant evo-devo. *Frontiers in plant science*, 4:321, 2013.

[59] J. Spaethe, A. Schmidt, A. Hickelsberger, and L. Chittka. Adaptation, constraint, and chance in the evolution of flower color and pollinator color vision. In *Cognitive ecology of pollination: animal behavior and floral evolution*. Cambridge University Press., 2001.

[60] J. Spaethe, J. Tautz, and L. Chittka. Visual constraints in foraging bumblebees: flower size and color affect search time and flight behavior. *Proceedings of the National Academy of Sciences*, 98(7):3898–3903, 2001.

[61] S. Stankowski, J. M. Sobel, and M. A. Streisfeld. The geography of divergence with gene flow facilitates multitrait adaptation and the evolution of pollinator isolation in Mimulus aurantiacus. *Evolution*, 69(12):3054–3068, 2015.

[62] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

[63] H. Stubbe et al. Genetik und Zytologie von Antirrhinum L. sect. Antirrhinum. 1966.

[64] C. Suchet, L. Dormont, B. Schatz, M. Giurfa, V. Simon, C. Raynaud, and J. Chave. Floral scent variation in two antirrhinum majus subspecies influences the choice of naïve bumblebees. *Behavioral Ecology and Sociobiology*, 65(5):1015–1027, 2011.

[65] J. Szymura. Analysis of hybrid zones with Bombina. *Hybrid zones and the evolutionary process*, pages 261–289, 1993.

[66] E. Tastard, C. Andalo, M. Burrus, L. Gigord, and C. Thébaud. Effects of floral diversity and pollinator behaviour on the persistence of hybrid zones between plants sharing pollinators. *Plant Ecology & Diversity*, 7(3):391–400, 2014.

[67] H. Tavares. *Evolutionary genetics and genomics of flower colour loci in an Antirrhinum hybrid zone*. PhD thesis, University of East Anglia, John Innes Centre, 2014.

[68] H. Tavares, A. Whibley, D. L. Field, D. Bradley, M. Couchman, L. Copsey, J. Elleouet, M. Burrus, C. Andalo, M. Li, et al. Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, 115(43):11006–11011, 2018.

[69] V. Vadez, J. Kholová, G. Hummel, U. Zhokhavets, S. Gupta, and C. T. Hash. Leasyscan: a novel concept combining 3d imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget. *Journal of Experimental Botany*, 66(18):5581–5593, 2015.

[70] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.

[71] M. Wheldale. The inheritance of flower colour in Antirrhinum majus. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 79(532):288–305, 1907.

[72] A. C. Whibley. *Molecular and genetic variation underlying the evolution of flower colour in Antirrhinum*. PhD thesis, University of East Anglia, 2014.

[73] A. C. Whibley, N. B. Langlade, C. Andalo, A. I. Hanna, A. Bangham, C. Thébaud, and E. Coen. Evolutionary paths underlying flower color variation in Antirrhinum. *Science*, 313(5789):963–966, 2006.

[74] H. M. Whitney, G. Milne, S. A. Rands, S. Vignolini, C. Martin, and B. J. Glover. The influence of pigmentation patterning on bumblebee foraging from flowers of Antirrhinum majus. *Naturwissenschaften*, 100(3):249–256, 2013.

[75] A. J. Wilson, D. Reale, M. N. Clements, M. M. Morrissey, E. Postma, C. A. Walling, L. E. Kruuk, and D. H. Nussey. An ecologist's guide to the animal model. *Journal of animal ecology*, 79(1):13–26, 2010.

[76] J. Wilson. The inheritance of milk yield in cattle. *Scientific Proceedings of the Royal Dublin Society*, 1911.

[77] A. Wollstein, S. Walsh, F. Liu, U. Chakravarthy, M. Rahu, J. H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J. R. Vingerling, et al. Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. *Scientific reports*, 7(1):1–11, 2017.

[78] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.