The repeatability of evolution; colour pattern control in *Heliconius* butterflies

Jake Morris

Doctor of Philosophy

University of York

Biology

December, 2016

Abstract

Heliconius butterflies are found across the neo-tropics, with bright aposematic colour patterns. These Müllerian mimics show striking colour pattern convergence across species, while paradoxically showing striking diversity within species. Thus *Heliconius* wing patterns have become an excellent system for understanding the repeatability of evolution. This work has identified a number of genes that appear to be involved in colour pattern control across species, such as *optix* and *cortex*, which respectively control red and yellow pattern elements. However, this work has only looked at the genetic basis of colour pattern in a small number of species, and primarily focuses on just two; *H. melpomene* and *H. erato*. I first use a population genomics approach to try to identify whether *optix* controls the hindwing rays phenotype in two poorly studied species; *H. demeter* and *H. aoede*. I identify both divergence associated with colour pattern at this *optix*, as well as another putative colour pattern control locus in *H. aoede*, the ommochrome pathway gene *cardinal.* Further, I use Quantitative trait loci analysis to explore the genetics of colour pattern in *H. melpomene,* confirming *WntA* as the gene controlling the 'broken band' phenotype and I identify a locus associated with red-orange pigmentation, while also exploring the role of minor effect loci in quantitative colour pattern variation. Finally, I use the natural diversity at two hybrid zones, in conjunction with phylogenetic discordance at mimicry loci, to identify putative regulatory enhancers associated with colour pattern shifts, investigate introgression across species at this fine genetic scale, and the possible role of colour pattern introgression in *Heliconius* speciation. This work reveals both interesting cases of convergent genetic evolution, independent genetic evolution and introgression, showing that a variety of evolutionary processes have shaped *Heliconius* mimicry across species.

Contents

List of tables

List of figures

Acknowledgments

I have many people that I owe a debt of gratitude to for helping me in completing this thesis. However, no one deserves more thanks than my supervisor Kanchon Dasmahapatra, in Kanchon I could have not asked for a better mentor. He has helped in innumerable ways, with experimental design in the early stages, library prep midway, and feedback and comments on drafts towards the end. Perhaps most of all though, he has been supportive throughout, even putting me up under his own roof for a period when needed. I have on the whole found this experience to be one of enjoyment and excitement, which I owe in large part to Kanchon, in particular, as his first PhD student I hope he has also found this experience to be both of these two things.

There are a number of other people that I must also thank for their intellectual input, help and support. Chief among them is Neil Rosser, in the early stages he was a housemate, and throughout has been a friend and colleague, discussing ideas with me when needed and helping supply butterflies and catch/find them in Suriname. I must also thank Chris Thomas and Peter Young for their role as TAP panel members, challenging, supporting and encouraging me throughout. I must also thank the other members of Kanchon's group, in particular Lucie Queste and Bruna Cama, as well as my wider departmental colleagues at York, especially Colin Beale for his thoughts on some statistical analyses.

As Kanchon's only student, and for a long period the only lab group member other than himself based in York, a number of people in Chris Jiggin's group in Cambridge have been very generous with their time, acting as surrogate colleagues. Primarily, and in no particular order, I have to thank Richard Merill for help with QTL analysis and other ideas; Richard Wallbank for thoughts on convergent evolution and *H. elevatus*

introgression; Simon Martin for his help with Twisst and thoughts on gene flow and introgression, as well as finding/catching butterflies in Suriname; the intimidatingly bright and incredibly helpful John Davey; and Pasi Rastas for his help with linkage mapping. I'd also like to thank the wider *Heliconius* community, especially Mathieu Joron for supplying butterflies and samples, and Jim Mallet, Owen McMillan and Chris Jiggins all of whom I have asked questions of at one time or another. Their depth of knowledge and willingness to help is inspiring.

My research was also helped and assisted by many others in more practical ways. Fieldwork in Suriname would not have been possible without the fantastic local knowledge of Ewout Eriks at the Neotropical Butterfly Park (Suriname). Thanks also to Paul Scott and Chris Lancaster and the rest of the horticultural team, without their expertise in looking after plants (*Passiflora*) and help with feeding butterflies this thesis would not have been possible. I owe thanks also to Florian Mann (at Technische Universität Braunschweig) for his work doing the GC-MS analysis of pheromones. Thanks also to Lauren Rawlins at York (Environment) who spent a summer helping me phenotype butterflies from the crosses. Stuart priest and Andrew Smith have also been of great help in computing matters, the former with Kanchon's server '*Heliconius*' and the latter with the N8 and YARCC clusters.

Finally, there are many people who, while they may have not helped with research, but who have been integral to this process (and my sanity) as friends here in York. Most of all though I have to thank my partner Tasha, who has put up with me and supported me throughout, even catching butterflies with me in Suriname. I could not have done any of this without her and I look forward to our next chapter together (now these ones are finally written). Last but not least, I'd like to thank my parents, without whose support (financially and emotionally) I could not have got here.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references. Where work that contributed to this thesis was undertaken by someone else this has been indicated and stated below:

Kanchon Dasmahapatra (The Univeristy of York) carried out both WGS and RAD library preparations.

Florian Mann (Technische Universität Braunschweig) carried out GC-MS analysis for the quantification and detection of chemicals in pheromone extracts.

A number of others at Cambridge University also contributed advice. Simon Martin advised on the analyses involving Twisst (and contributed publically available scripts). Pasi Rastas and John Davey (who contributed a personal script) advised on Linkage mapping, while Richard Merrill and Nicolas Navarro (Université Bourgogne Franche-Comté) advised on the QTL analysis.

1. *Heliconius*: colour pattern diversity and convergence

Estimates of the number of species on earth are constantly being revised and updated. Around 1.5 million eukaryotes are known to science and named (Costello *et al.* 2013), but another 3.5 – 7.2 million, depending on who you ask, are as yet unknown (Mora *et al.* 2011; Costello *et al.* 2013). This is just a fraction of the estimates for microbial species that suggest there could be anywhere upward of 1 trillion different species (Locey & Lennon 2016). What all these estimates underline is that the world is hugely diverse. This diversity is the result of several billion years of evolution. Progressing our understanding of the complex evolutionary processes that have generated this biodiversity is more important than ever, given the future of unprecedented, global environmental change that we face.

1.1.1 Evolution: progress and paradigms

In 'The Structure of Scientific Revolutions' (1970), Thomas Kuhn laid out a model for how science progresses. He argued that science in any given field progresses and evolves through a process of punctuated-equilibrium, in which long periods of puzzle-solving 'normal science' in which researchers fit observations and empirical work into an existing framework, are punctuated by periods of uncertainty and crisis. These crises arise through an accumulation of observations that don't fit into the current theoretical framework, or as Kuhn coined, the 'paradigm', thus undermining and destabilising the field and leading to the rise of a new and better paradigm that can explain the old data as well

as the new. Kuhn proposed that it is through these paradigm shifts that science makes its great leaps in understanding.

Our understanding of evolutionary biology has though arguably progressed in a somewhat different fashion, with perhaps the last paradigm shift extending back to the modern synthesis and the birth of Neo-Darwinism (Huxley 1942). However, although a very real leap in understanding, even this still very much built upon the ideas in Darwin's 'Origin of Species' (1859) rather than being a wholesale replacement of them. Rather than wholesale changes to paradigms, our ever increasing understanding of evolutionary biology has led to smaller shifts that tweak or extend the current framework, driven in large part by technological advancements and concerted efforts studying a small number of model species (Pigliucci 2007).

1.1.2 **Rapid improvements in technology**

In 1944, Avery *et al* published their seminal paper in which they suggested that DNA, not proteins as had previously been thought, was the carrier of genetic information. Less than a decade later Watson and Crick published the complete structure of this most important of molecules (Watson & Crick 1953). Only ten or so years later than this, and the genetic code was finally being cracked with the help of Marshall W. Nirenberg and his team (Nirenberg & Matthaei 1961; Matthaei *et al.* 1962; Kellogg *et al.* 1966). This sudden burst of interest and discovery from the 1940s to the 1960s, took science from simply understanding the importance of nucleic acids, to understanding their structure and the basics of how they pass on the genetic information they encode. Armed with this knowledge, the technology to exploit this new frontier in biology quickly followed, with the isolation of Restriction enzymes (Smith & Welcox, K 1970), the development of Sanger sequencing in the mid-1970s (Sanger & Coulson 1975), and the development of the Polymerase Chain Reaction in the 1980s (Saiki *et al.* 1985; Mullis & Faloona 1987). In the last few decades, technology has progressed not so much steadily as exponentially, in

line with Moore's law, with the cost of sequencing per base continually falling (Pettersson *et al.* 2009). We have progressed from one sequenced human genome in 2001 (Lander *et al.* 2001), completed at a cost of more than $3 billion dollars over 15 years or so (Shendure *et al.* 2004), to over 179 individually sequenced genomes less than a decade later (Altshuler *et al.* 2010). This rapid development of new and improved technologies has opened up both new fields to research, as well as a vastly greater array of taxa.

1.1.3 **Expanding model species**

Evolutionary research was for a long time dominated by concerted efforts to understand just a small fraction of the planet's species. By concentrating efforts on such a small number of species, new research was able to build upon and add to a larger body of past research. However, the choice of organism has often been as much the result of historical accident as of reasoned choice (Powell 1997). The most famous model organisms have been the fruit fly *Drosophila melanogaster,* and the mouse *Mus musculus* (Hedges 2002). Through these species our understanding of genetics and species development has advanced greatly. However, focussing our attention on such a small number of model species inherently limits the range and complexity of evolutionary processes that can be understood.

Over the past several decades, the advent of genetic sequencing has led to a broadening of the definition of model species from just *Drosophila melanogaster*, *Mus musculus* and a handful of others (Hedges 2002), to now include a selection of other less tractable but more ecologically variable species and systems in which evolutionary and genetic research can be carried out (Mallet 2006; Ellegren 2014). Today in the era of next generation sequencing, even relatively small efforts or single laboratory groups can build reference-quality genomes for their study species of choice, and can sequence genomic markers from 100s of individuals more (Ellegren 2014). This opening up of species and biological systems to research, through technological advances, has itself also contributed to

extending and shifting our understanding of a whole variety of evolutionary processes that both generate biodiversity and determine how evolution progresses. This has had particular impact on our understanding of the repeatability of evolution (Elmer & Meyer 2011).

1.2 CONVERGENT EVOLUTION

Richard Owen in his work 'On the Archetype and Homologies of the Vertebrate Skeleton' (1848) defined an analogous structure as 'a part or organ in an animal that has the same function as another part or organ in a different animal'. Owen used the example of flying dragons (*Draco volans*) and birds to illustrate the difference between analogous structures like the wings or parachutes they both use and others that are homologous like each of their forelimbs. This important distinction allowed science to see for the first time that while some species were similar due to sharing the same features, others were similar despite their obvious differences. Today in an evolutionary context, we can use this concept to see how different species have independently evolved similar traits, as adaptations to similar environmental pressures and life histories, which allow them to survive and thrive, a phenomenon termed 'convergent evolution'. Convergent evolution can occur at two levels: the first level is the phenotype, as Richard Owen noted in the evolution of birds, bats and pterosaurs; and the second level is the genotype, characterised by species that may or may not be closely related, but have evolved similar traits through changes to the same genetic and developmental pathways.

1.2.1 **Convergent genetic evolution**

Convergent evolution can occur in two ways. Firstly, through mutations that affect *different* genetic and developmental pathways; this is illustrated in the depigmentation between some Mexican cave tetra populations, which can be caused by mutations at both the gene *Oca2*, or at the gene *Mc1r* (Gross *et al.* 2009). Alternatively, mutations can arise

independently but affect the same genes in the same genetic and developmental pathways (Zhang 2006; Tishkoff *et al.* 2007; Gompel & Prud'homme 2009; Parker *et al.* 2013). We can now identify many examples of convergent genetic evolution between taxa at a range of different taxonomic levels and for a wide range of different traits, from the evolution of echolocation between distantly related bats and dolphins (Parker *et al.* 2013), to digestive system efficiency between Asian (*Pygathrix nemaeus*) and African (*Colobus guereza*) colobine monkeys (Zhang 2006), to lactase persistence in multiple human populations (Tishkoff *et al.* 2007). Perhaps one of the most striking and pervasive examples of convergent evolution is that of the gene *Melanocortin-1 receptor* (*Mc1r*). This has been found to be the root cause of a surprisingly large number of pigmentation changes across vertebrate taxa, including a number of birds (Mundy 2005), fish (Gross *et al.* 2009) and mammals (Eizirik *et al.* 2003; Hoekstra *et al.* 2006; Dun *et al.* 2007). This recurrent deployment of Mc1r naturally leads on to questions asking why this kind of convergent genetic evolution occurs, and what properties do certain genes have, that might cause their continued usage across taxa.

In effect these are examples of evolutionary repeatability. Understanding just how often evolution repeatedly solves the same problems with the same solutions, can inform us of the kinds of constraints placed upon evolution. When Stephen J. Gould (1990) wrote of replaying the 'tape of life', he proposed that evolution would take a very different path each time, due to its inherently stochastic nature. However, with the number of genomic datasets increasingly becoming available in diverse taxa, the genetic basis of many traits is becoming better understood and the frequency of convergent genetic evolution is also becoming better understood (Stern 2013). It is now clear that in a large proportion of cases, estimated at 0.32 on average, when similar traits evolve independently, the root cause is often mutations at the same genes and loci (Conte *et al.* 2012).

While we have many examples that allow us to better understand the frequency of convergent evolution, our understanding of why convergence occurs is still based principally on a smaller number of semi-model systems that have been extensively studied. These studies have begun to reveal a number of important properties that might cause repeated evolution and use of the same gene across taxa. However, while all of selection, mutation, recombination, pleiotropy, epistasis, and developmental and genetic architecture can lead to repeated evolution, it is still unclear whether some of these properties, and which ones, play a primary role more frequently in the repeated convergent genetic evolution of a phenotype. The reason for this is that unpicking the various contributions and effects of any of these properties on any one trait is difficult at best (Gompel & Prud'homme 2009; Stern & Orgogozo 2009; Barrett & Hoekstra 2011); while all can contribute to whether a gene or locus may be more likely to be involved in convergent evolution (Gompel & Prud'homme 2009; Stern 2013). For example, theoretically a gene might be more likely to evolve beneficial mutations (Orr 2005a), if the epistatic background on which a genetic change finds itself promotes its advantageous evolution amongst closely related taxa; while in contrast, a lack of negative pleiotropic genetic effects might also free it of the evolutionary shackles perhaps constraining other evolutionary paths (Weinreich 2006).

1.2.2 **Parallel evolution: a special case of convergence**

The distinction between convergent evolution and parallel evolution has at times been both confusing and controversial, and some have declared it a false dichotomy entirely (Arendt & Reznick 2008). Before the advent of modern genetics, convergent evolution was understood as convergence between unrelated species (Stern 2013), and was often assumed to have been caused by quite different changes (Arendt & Reznick 2008), which we would today describe as changes in different genetic and developmental pathways. In contrast, parallel evolution was the term used to describe phenotypic changes between

related species (Stern 2013). Today, all definitions of parallel evolution understand it as involving convergent phenotypic changes that deploy the same genetic and developmental pathways in their control (Arendt & Reznick 2008; Stern & Orgogozo 2009; Conte *et al.* 2012; Stern 2013).

In this thesis, I use parallel evolution to describe only situations in which similar phenotypes have been arrived at, in different species or populations, through changes to the same genetic and developmental pathways, and most importantly from similar genetic starting points, i.e. from populations that are likely not completely reproductively isolated and can still hybridise. Examples include the parallel genetic evolution of pelvic and armour reduction in multiple freshwater stickleback populations (Cresko *et al.* 2004; Shapiro *et al.* 2004), and in Mexican cave tetra's where the gene *Oca2* has independently led to the evolution of depigmentation in different populations (Protas *et al.* 2006). Hence parallel evolution becomes a special case of convergent evolution.

1.2.3 **Introgression**

Convergent evolution may be driven by convergent genetic evolution, or alternatively by 'collateral evolution' (Stern 2013), which occurs through the shared presence of alleles among populations or species, either through shared ancestry and incomplete lineage sorting, or through the introgression of these alleles from one population/species to another (Stern 2013). Introgression occurs through hybridization, when individuals from different species crossbreed and exchange genetic information across this species boundary (Twyford & Ennos 2012). Introgression has the power to be a creative evolutionary process, allowing advantageous alleles and adaptive allelic combinations to accumulate faster than by mutation alone, and can even potentially drive speciation, if the traits that introgress are involved in mate choice, or lead to reduced hybrid fitness in either environment of the parental species (Gompert *et al.* 2006). However, evidence of adaptive introgression is rare, with examples known only from a small number of diverse

taxa, such as insecticide resistance in *Anopholes* mosquitoes and poison resistance in mice (Song *et al.* 2011); in the tunicate *Ciona intestinalis* (Roux *et al.* 2013); between ancient humans Neanderthals and Denisovans (Racimo *et al.* 2015); and in Darwin's finches (Lamichhaney *et al.* 2015), as well as a handful of other potential examples (Hedrick 2013). These examples are few and far between, and in some cases more suggestive than certain. In addition in many of these cases, such as those of mice, mosquitoes and coyote and wolf coat colour (Hedrick 2013), human disturbance and influence is implicated in driving or assisting this process.

1.3 THE *HELICONIUS* RADIATION

1.3.1 **Ecology and mimicry**

Heliconius butterflies are a genus of about 40 species found across the neo-tropics. These butterflies have strong chemical defences, through cyanogenic glycosides both sequestered from their *passiflora* host plants and synthesised *de novo* (Engler-Chaouat & Gilbert 2007; Hay-Roe & Nation 2007). This toxicity has led to the evolution of bright aposematic colour patterns, with nearly, but not quite all species Müllerian mimics (Merrill *et al.* 2012). These show colour pattern convergence across species, while also paradoxically showing striking diversity within species as different colour patterns have become the optimal pattern in different geographic areas (Joron & Mallet 1998). This divergence within species leads to numerous intraspecific hybrid zones across the neotropics at which recombinant forms between geographic colour pattern races are found (Mallet 1986; Blum 2008). However, the integrity of these races are maintained by strong frequency-dependent selection against these non-mimetic recombinant forms, imposed by predation against these rare phenotypes (Mallet & Barton 1989b; Sherratt 2006; Merrill *et al.* 2012). This diversity (Figure 1.1) has made *Heliconius* an important and tractable system for the study of convergent evolution.

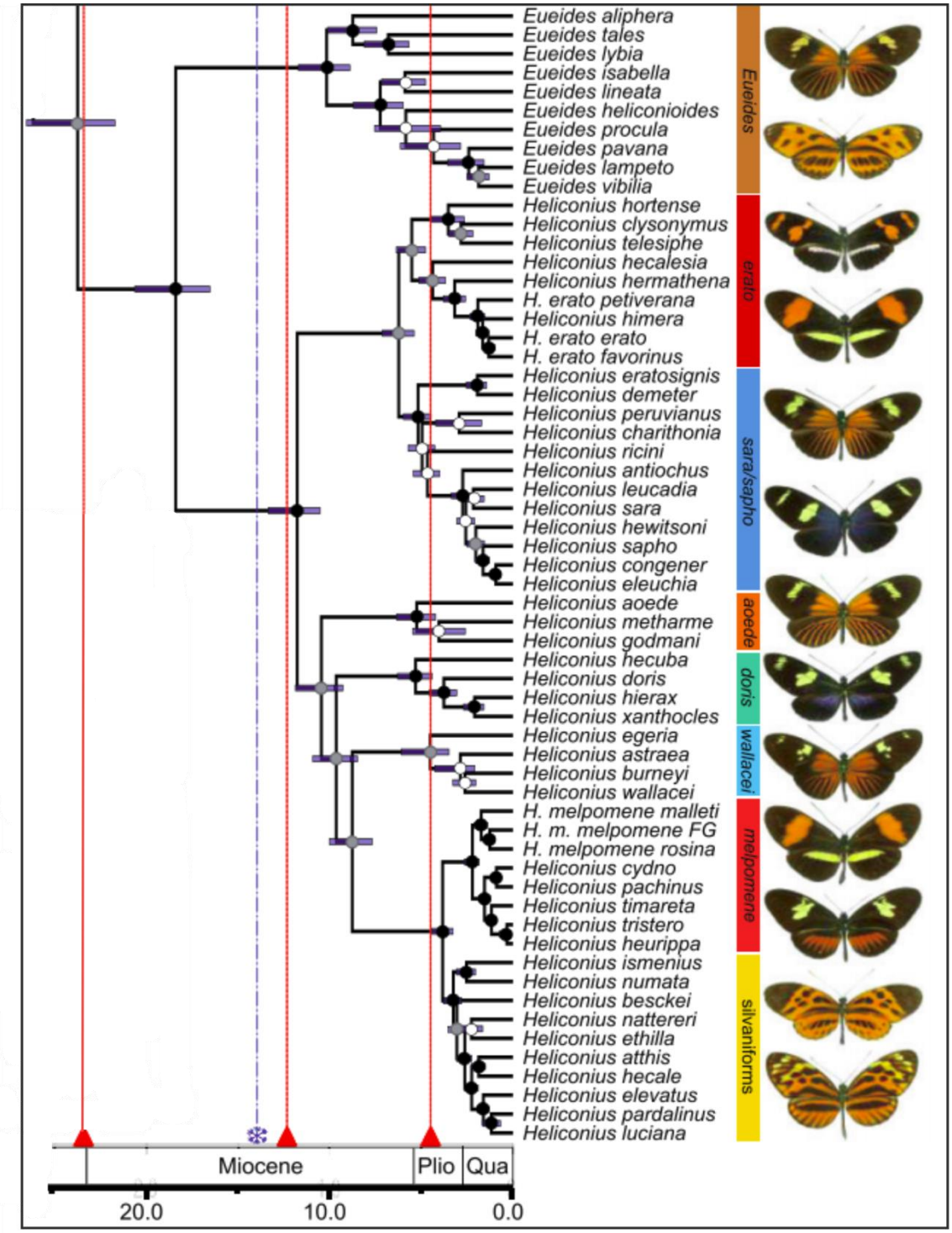


Figure 1.1 - Dated Multi-locus species tree of *Eueides* and *Heliconius* species from Kozak *et al.* (2015), estimated using 20 nuclear and 2 mitochondrial markers with an uncorrelated molecular clock method (BEAST). Bars signify the 95% credible intervals around the mean node ages. Scale axis in Ma. Deep splits are shown within the well-studied *H. erato* and *H. melpomene*. Figure also adapted from Kozak *et al.* (2015), and shows important taxa used in this thesis from the silvanifrom, *H. melpomene*, *H. erato*, *sara/sapho* (with *H. demeter*) and *H. aoede* clades.

1.3.2 **The Nymphalid groundplan**

Studies of *Heliconius* wing patterns do not exist in a vacuum. Research across a wide range of Lepidoptera taxa has deepened our understanding of the evolution of colour pattern diversity. In order to fully understand convergent evolution, examples like this, in which phenotypic diversity is understood at all levels, are important This involves understanding how selection varies across the genome; the function of genes under selection that drive diversity; and the developmental pathways that gene selections are part of (Brakefield *et al.* 1996; Brakefield 1998). Few other phenotypes found across such a diverse range of taxa are as well understood as butterfly wing patterns (Brakefield 1998), with much of this work having focussed on *Heliconius* species, as well as the Forest Brown butterfly *Bicyclus anynana,* and the Buckeye butterfly *Precis coenia*.

Nymphalid colour pattern is hypothesised to be determined by an underlying nymphalid groundplan (Nijhout 1990). Under this framework, butterfly wing pattern elements, specifically bands, chevrons and eyespots, can be explained by a number of separate independent symmetry systems (Nijhout 1994). The wing is then further compartmentalised into cells by wing venation, with each cell independent of others and containing one or several elements from each symmetry system (McMillan *et al.* 2002). This modular system, means that regulatory changes to genes can affect one part of the pattern while leaving others unaffected (Beldade & Brakefield 2002), giving enormous flexibility and resulting in the vast array of patterns seen in extant species today.

Much of the pioneering work into understanding the nymphalid groundplan focused on the evolution of butterfly eyespots (Brunetti *et al.* 2001). This work shows that pigments for each eyespot are deposited around a central point or 'focus', normally found midway between wing veins (McMillan *et al.* 2002). These focus points signal to other surrounding cells, that then differentiate into the diversity of pigment scale types that pattern the eyespot. Experimental manipulation, using both transplants and damage to focus points

(Brakefield *et al.* 1996) in combination with artificial selection experiments, have further revealed that different aspects of eyespot patterning, for example size and colouration, are uncoupled and independent (Monteiro *et al.* 1997; Brunetti *et al.* 2001). These findings were the first to indicate both the flexibility and modular nature of these traits that have led to the diversity of butterflies.

1.3.3 Convergent genetics of *Heliconius* wing patterns

Understanding the genetics of colour pattern diversity in *Heliconius* began well before the advent of modern DNA sequencing, with the work of Turner and Crane (1962) and Sheppard (1985). These studies utilised the diversity of colour patterns found within species to conduct crossing experiments studying the segregation of different colour pattern elements, and therefore to infer the number of loci controlling them. This work identified a small number of major effect loci that act as colour pattern switches, to explain much of the diversity in red-orange and yellow pattern elements. Subsequent work using QTL (Quantitative trait locus) mapping analyses in the two species *H. melpomene* and *H. erato*, as well as a small number of others, has shown that the loci controlling mimicry in *Heliconius* are the same in both species (Baxter *et al.* 2008a; Papa *et al.* 2013; Nadeau *et al.* 2014; Huber *et al.* 2015). This confirms that colour pattern diversity in *Heliconius* has evolved through convergent evolution.

The most well understood of these major effect loci is the *BD* locus, that controls the main red-orange elements in *H. melpomene* and *H. erato* (Sheppard *et al.* 1985). This locus has been mapped to chromosome 18 in both species (Baxter *et al.* 2008b; Papa *et al.* 2013), as well as in *H. hecale* (Huber *et al.* 2015). Furthermore, the gene controlling these colour pattern elements has been shown to be the transcription factor *optix*. This gene has been found to be expressed during development in red regions of the wing just prior to ommochrome pigmentation (Reed *et al.* 2011). This expression pattern has been

confirmed in both *H. erato* and *H. melpomene* as well as a range of other *Heliconius* species quite closely related to the latter (Reed *et al.* 2011; Martin *et al.* 2014b).

The *Yb/N,* locus has for a long time been shown to control both the hindwing yellow bar, and together with the *BD,* the switch from a red to yellow forewing band in *H. melpomene* and *H. erato* (Sheppard *et al.* 1985). QTL mapping studies have now identified this locus in both species on Chromosome 15 (Baxter *et al.* 2008b; Papa *et al.* 2013). In other *Heliconius* this locus has also been implicated in controlling a number of other yellow-black patterning traits, such as yellow/white hindwing margins and patterns across the apical part of the forewing (Linares 1996; Jiggins & McMillan 1997; Jiggins *et al.* 2005; Kronforst *et al.* 2006a; Ferguson *et al.* 2010; Huber *et al.* 2015). This locus has also been found to be homologous with the supergene *P,* which alone controls all patterning in the polymorphic species *H. numata* (Joron *et al.* 2006b; Jones *et al.* 2012). A candidate gene for controlling this colour pattern diversity has also been identified: *cortex*, which has been found to be a divergent between races and species of *Heliconius* with different phenotypes, as well as showing differential gene expression between black and yellow wing regions during development (Nadeau *et al.* 2016). Furthermore this gene is not just implicated in colour patterning in *Heliconius* species, but also in the silk moth *Bombyx mori* and the peppered moth *Biston betularia* and *Bicyclus anyanna* (Nadeau *et al.* 2016; Hof *et al.* 2016), a true hotspot for convergent evolution.

Another mimicry locus has also been identified on chromosome 10. This third major wing colour pattern locus contributes to both forewing band shape and melanisation across the discal portion of the forewing (Kronforst & Papa 2015). Again, at this locus, a candidate gene has been identified; *WntA* (Martin *et al.* 2012; Gallant *et al.* 2014a; Kronforst & Papa 2015). This is supported both by QTL mapping studies in *H. cydno* and *H. erato* (Martin *et al.* 2012; Papa *et al.* 2013) as well as for band shape in *H. melpomene* (Martin *et al.* 2012). This QTL analysis has been backed up by *in situ* hybridisations that show *WntA* expression

concordant with melanic scales during development (Martin *et al.* 2012; Gallant *et al.* 2014a). Furthermore, *WntA* has been found to control melanic patterning in *Limenitis arthemis*, a species highly divergent (>65 million years) from *Heliconius*. Altogether this body of work reveals stunning convergence in the genetic evolution of wing patterns, not just within *Heliconius* but in some cases across highly divergent species.

1.3.4 **Modulation of mimicry**

Genes involved in butterfly eyespot development have been found to be part of ancient developmental pathways, which are proposed to have either been co-opted for the control of butterfly wing patterning or built de-novo using the same genes (Monteiro & Podlaha 2009). Butterfly eyespots have been found to be controlled by a variety of genes, including *Distal-less* and *spalt* (Beldade *et al.* 2002; Monteiro *et al.* 2013; Zhang & Reed 2016). These two genes play different roles, with *spalt* promoting eyespot formation and *Distal-less* repressing eyespot development (Zhang & Reed 2016). However, both also play roles in insect appendage patterning, with *spalt* expressed in antenna during development and *Distal-less* found to specify insect limbs and beetle horns (Monteiro & Podlaha 2009). The genes involved in *Heliconius* wing patterning have also been shown to have conserved developmental functions across taxa.

The gene *cortex* is involved in cell-cycle regulation across species, and is therefore predicted to control colour patterning through controlling scale cell development (Nadeau *et al.* 2016), while *WntA* is a conserved morphogen from the Wnt family of signalling molecules which includes the gene *wingless*. This gene, *wingless* has been shown to be involved in wing pigment patterning in both *Drosophila* (Swarup & Verheyen 2012) and other Lepidoptera (Martin & Reed 2010), and is also at the genomic locus that controls the white/yellow switch in some *Heliconius* (Kronforst *et al.* 2006b). Furthermore, while *optix* has not yet been found to control wing patterning in species outside of *Heliconius*, it does appear to play a conserved role in the determination of

scales that link the forewing and hindwing together, so again this gene appears to have an ancient and quite possibly conserved role in wing development (Reed *et al.* 2011; Martin *et al.* 2014b). It seems likely in these cases that these genes and developmental networks have been co-opted to control wing patterning, with different ommochrome pigments and in different parts of the wing.

This kind of co-option cannot be achieved through changes in protein coding sequence, which through most of the twentieth century was thought to be primarily responsible for phenotypic evolution. However, such regulatory changes can be achieved through evolution at *cis*-regulatory modules which can rapidly drive morphological evolution (Wray 2007; Wittkopp & Kalay 2012). The two most well understood forms of *cis*-regulatory sequences are promotors and enhancers, with enhancers the main driving force behind rapid morphological evolution (Wittkopp & Kalay 2012). A single gene can have multiple enhancers, with each controlling the expression of a gene in a different cell type or at time of development. In this way genes can be involved in multiple developmental networks, and through changes in an enhancer the function of that gene can be conserved across other networks (Monteiro & Podlaha 2009). A certain amount of evolutionary stability is also built into this system of modulation, with multiple enhancers having overlapping functions (Hong *et al.* 2008; Cannavò *et al.* 2016). This modulation provides a flexible toolkit through which gene expression changes can rapidly alter phenotypes and drive adaptive evolution (Wray 2007).

This enhancer modulation has been identified in a number of cases of parallel and convergent evolution, with deletion or mutations at a single enhancer having major adaptive phenotypic effects (Stern 2013). Two of the most well characterised examples of convergent evolution through changes at regulatory enhancers are the *shavenbaby* gene in *Drosophila* species, and pelvic reduction in sticklebacks through deletion of the *Pitx1* gene (McGregor *et al.* 2007; Chan *et al.* 2010; Frankel *et al.* 2012). Furthermore, in *H.*

melpomene and related species, two putative regulatory modules have now been identified around the gene *optix*, with each thought to control a different pattern element (Wallbank *et al.* 2016). This work took advantage of shared ancestry between species from the *H. melpomene* clade to identify these modules, as adaptive introgression of these colour pattern loci appears to have driven mimicry between these species through the shuffling of regulatory enhancers (Dasmahapatra *et al.* 2012; Pardo-Diaz *et al.* 2015; Wallbank *et al.* 2016).

1.3.5 *Heliconius* mimicry and speciation with gene flow

While adaptation is generally thought of at the unit of the gene, reproductive isolation has often been viewed as occurring between whole genomes. However, the genic view of speciation (Wu 2001) highlights that reproductive isolation can be achieved by differentiation at just a small number of loci that have a disproportionate effect on divergence between populations. With the advent of new genome, and genomic marker, sequencing methods, the field of speciation genomics was born (Feder *et al.* 2012). These novel methods have now been used to identify the number and types of loci involved in species differentiation in a wide variety of species, from *Timema* stick insects (Soria-Carrasco *et al.* 2014) and *Chorthippus* grasshoppers (Berdan *et al.* 2015), to the *Drosophila simulans* clade that Wu worked on (Garrigan *et al.* 2012), to Lord Howe Island palms (Savolainen *et al.* 2006), and Cichlid species (Mattersdorfer *et al.* 2012; Keller *et al.* 2013).

These methods have also been used to reveal the signatures and effects of heterogeneous genetic divergence on the genome. One such effect is hitchhiking, which leads to linked regions of the genome around loci showing elevated divergence due to a reduced effective migration rate (Via & West 2008). It is proposed that if enough regions of the genome exhibit divergence hitchhiking this can then lead to almost complete genome hitchhiking, in which the effective migration rate and therefore divergence is

reduced genome wide (Feder *et al.* 2012). Across intraspecific hybrid zones in *Heliconius* colour pattern loci can be identified as clear 'islands of divergence' (Baxter *et al.* 2010; Counterman *et al.* 2010; Nadeau *et al.* 2013, 2014). These stand out, as strong selection on these important adaptive loci reduces their gene flow relative to the rest of the genome where neutral or similarly adaptive loci are free to be shared (Wu 2001). One of the important outcomes from these hitchhiking effects, is to considerably increase the likelihood of speciation in the face of gene flow (Feder *et al.* 2012; Via 2012).

As well as playing an important role in mimicry, colour pattern has also been found to be used as cues for mate choice and species recognition (Merrill *et al.* 2011, 2014). This therefore makes colour pattern in *Heliconius* a 'magic trait' (Gavrilets 2004), with strong disruptive ecological selection also having the potential to drive assortative mating and reproductive isolation (Servedio *et al.* 2011). 'Magic traits' can greatly enhance the likelihood of speciation with gene flow, as recombination is prevented from dissociating these two important traits' effects (Servedio *et al.* 2011). While 'magic traits' are known from other taxa (Summers *et al.* 1999; Boughman *et al.* 2005; Hendry *et al.* 2009), the *Heliconius* example is one of the most well studied, and makes *Heliconius* an excellent system for the study of speciation with gene flow. All the same this work has mostly been on just two species, *H. melpomene* and *H. cydno* (Naisbit *et al.* 2001; Merrill *et al.* 2011, 2014), while the roles that other traits, like pheromones, might play have been largely overlooked (Jiggins 2008).

1.4 CONCLUSION

Overall, *Heliconius* wing patterns are an excellent system for understanding the repeatability of evolution, with ancient developmental genes likely having been co-opted in *Heliconius* into new patterning pathways that control mimicry. This appears to have been driven both by convergent genetic evolution between more divergent species, and

adaptive introgression between more closely related species. While *Heliconius* is already a model system for exploring questions around convergent evolution, next generation sequencing technology is now available to further our understanding of the convergent genetic basis of mimetic *Heliconius* wing colour patterning. This can be done by expanding the species, races and loci investigated. In the first three chapters of this thesis, I use a number of different approaches to do just that, and to investigate the repeatability in the evolution of *Heliconius* mimicry.

In chapter 1, I use a population genomics approach to try to identify whether the *BD* locus, a hotspot of evolution, appears to also control the common hindwing rays phenotype in species in which the genetic basis of mimicry has previously not been studied. In doing so I expand the range of taxa investigated, looking at species more distantly related from either *H. melpomene* or *H. erato* than have previously been studied. In chapter 2, I employ a different approach using QTL mapping analysis to confirm that the 'broken band' phenotype in *H. melpomene* is controlled by *WntA*, while also identifying a new locus associated with the shift from red to orange pigmentation, as well as other putative minor effect loci. In chapter 3, I take advantage of phylogenetic discordance at mimicry loci and natural diversity at two hybrid zones, to identify putative new enhancers associated with colour pattern shifts, and to refine other enhancers already identified across all three major mimicry loci. In doing so, I also investigate introgression across species at this finer level of genetic control. In the final chapter, I investigate the role of this introgression in the speciation of two sympatric sister species; *H. elevatus* and *H. pardalinus*. In these species, colour pattern loci can be clearly identified as 'islands of divergence' (Kryvokhyzha 2014) suggesting gene flow has and possibly still does occur across most of their genomes. I investigate both the role colour pattern preference and pheromones, in the speciation of these two sympatric species of *Heliconius*.

2. Population genomics of a multi-species mimetic hybrid zone

2.1 INTRODUCTION

Convergent evolution occurs when two or more species respond to selective pressures through the evolution of similar traits. In the past this was thought to predominantly occur through divergent genetic mechanisms and pathways (Stern 2013). However, it is now clear that a large proportion of the time, when similar traits evolve independently in different species or populations, the same genes and loci are often the cause of these convergent changes (Conte *et al.* 2012). Numerous examples of this have now been described, from the evolution of echolocation between distantly related bats and dolphins (Parker *et al.* 2013), digestive system efficiency between Asian (*Pygathrix nemaeus*) and African (*Colobus guereza*) colobine monkeys (Zhang 2006), to lactase persistence in multiple human populations (Tishkoff *et al.* 2007) and pelvic and armour reduction in multiple freshwater stickleback populations (Cresko *et al.* 2004; Shapiro *et al.* 2004). These examples show that convergent genetic evolution can occur between taxa, at many different levels of divergence.

The reasons as to why certain genes appear to be involved in repeated convergent evolution are often unclear. However, all of selection, mutation, recombination, pleiotropy, epistasis, or developmental and genetic architecture can contribute to whether a gene or locus may be more likely to be involved in convergent evolution

(Gompel & Prud'homme 2009; Stern 2013). Perhaps one of the most well-known examples of convergent evolution amongst distantly related taxa is the gene *Mc1r*. This has been implicated in changes in pigmentation in many vertebrate species, from mammals (Valverde *et al.* 1995; Eizirik *et al.* 2003; Römpler *et al.* 2006; Hoekstra *et al.* 2006; Dun *et al.* 2007) to birds (Mundy 2005) and even fish (Gross *et al.* 2009). This gene's role in pigmentation across such a broad array of taxa seems at first remarkable, but can in fact be at least partly explained by the conserved melanism pathway across vertebrates and the low pleiotropic effects *Mc1r*.

The growth in genetic and now genomic datasets has also led to a wealth of examples of parallel evolution, where convergence occurs between phylogenetically more closely related taxa. In *Astyanax mexicanus* (Mexican cave tetra), multiple populations have independently adaptated to cave-related conditions. Across different caves in Mexico populations show loss of pigmentation and regressed eyes among other novel features (Protas *et al.* 2007). These cave populations have evolved from a river, surface dwelling morph that can still interbreed with the cave forms. Quantitative Trait Loci analysis has shown that in three of these cave populations (Molino, Pachón and the inter-connected Yerbaniz and Japonés) albinism has been caused through adapted changes that lead to a loss of function in the protein *Oca2* (Protas *et al.* 2006).

The example of Mexican cave tetras perfectly demonstrates how similar adaptive pressures can cause lead to repeated phenotypic evolution driven by similar mutational changes (Protas *et al.* 2007). However, even in this example it has been found that genetic evolution is not always convergent. In some of these same cave systems (Pachón and Yerbaniz) as well as in several others (Curva, Piedras, Chica and Sabinos) another morph with reduced pigmentation can also be found. However, in this case it is not full albinism as caused by *Oca2* but instead changes in melanophore size controlled by *Mc1r* (Gross *et al.* 2009). Mexican cave tetras therefore show both how the same genetic pathway can

independently lead to the evolution of similar phenotypes in different populations, and how two similar phenotypes can be caused by mutations at two different genes in different populations. This is a perfect example of how evolution can come up with both convergent and different evolutionary answers to similar selective problems.

Heliconius butterflies are probably the most notable example of Müllerian mimicry found in nature. Found across South and Central America they show repeated phenotypic convergence across multiple species (Merrill *et al.* 2015). In addition to this convergence, natural intra-specific hybrid zones are found between neighbouring colour pattern forms (Mallet & Barton 1989a; Rosser *et al.* 2014). These narrow zones are maintained by strong frequency dependent selection against hybrid colour pattern form migration, as they match neither of the local optimal patterns on either side (Mallet 1986; Mallet & Barton 1989b). Wing pattern evolution has now been researched for over 50 years in *Heliconius*, and in a few species the loci controlling many aspects of these patterns have now been mapped to what is a relatively small number of loci of large effect (Sheppard 1963; Sheppard *et al.* 1985; Mallet 1989; Jiggins *et al.* 2005; Baxter *et al.* 2008b; Papa *et al.* 2013).

H. erato and *H. melpomene* show extraordinary diversity across South and Central America. Studies have therefore, for the most part focussed on these two distantly related species. Mimicry between these two species has been shown to have often evolved through convergent evolution at many of the same colour pattern loci (Baxter *et al.* 2008b). Even more strikingly, it has become clear that in some cases this is through regulatory changes using the same genetic pathways and genes. For example, the gene *WntA* is involved in the control of melanic patterning not only in some races of *Heliconius cydno*, *H. erato* and *H. melpomene* but also in other *Lepidoptera* species such as the American white admiral, *Limenitis arthemis* (Martin *et al.* 2012; Gallant *et al.* 2014a). Even more remarkable than this example, is that of the gene *cortex*. This has been found to be

at the centre of the *Yb* locus, which controls much of the yellow patterning in *Heliconius*. This gene has been found to not only control mimicry in several *Heliconius species* (*H. erato*, *H. melpomene*, *H. timareta* and *H. numata*) but to also be involved in controlling wing spots in another nymphalid, *Bicyclus anynana*, and in several species of moths *Biston betularia* and *Bombyx mori*, where it plays a role in the control of melanism (Nadeau *et al.* 2016).

Red patterning in *H. erato* and *H. melpomene* has also been mapped to a homologous genomic locus in both species (Baxter *et al.* 2008b). This locus has now been shown to contain the transcription factor *optix*, which is expressed in conjunction with red patterning in both *H. erato, H. melpomene,* and its sister species *H. cydno*, as well as in other *Heliconius* like *H. doris* and *H. atthis* (Reed *et al.* 2011; Martin *et al.* 2014b). In *H. melpomene,* and it's close relatives *H. timareta*, *H. cydno* and *H. elevatus* this locus has so far been further narrowed down into two regulatory subunits. These subunits are associated with the presence or absence of particular red colour pattern elements, either *dennis* or *rays* (Wallbank *et al.* 2016). These modules show distinct evolutionary histories within the wider *H. melpomene*/silvaniform clade. However, for both, interspecific introgression appears to have played a role through 'enhancer shuffling', with recombination between different species leading to new combinations and new diversity (Wallbank *et al.* 2016).

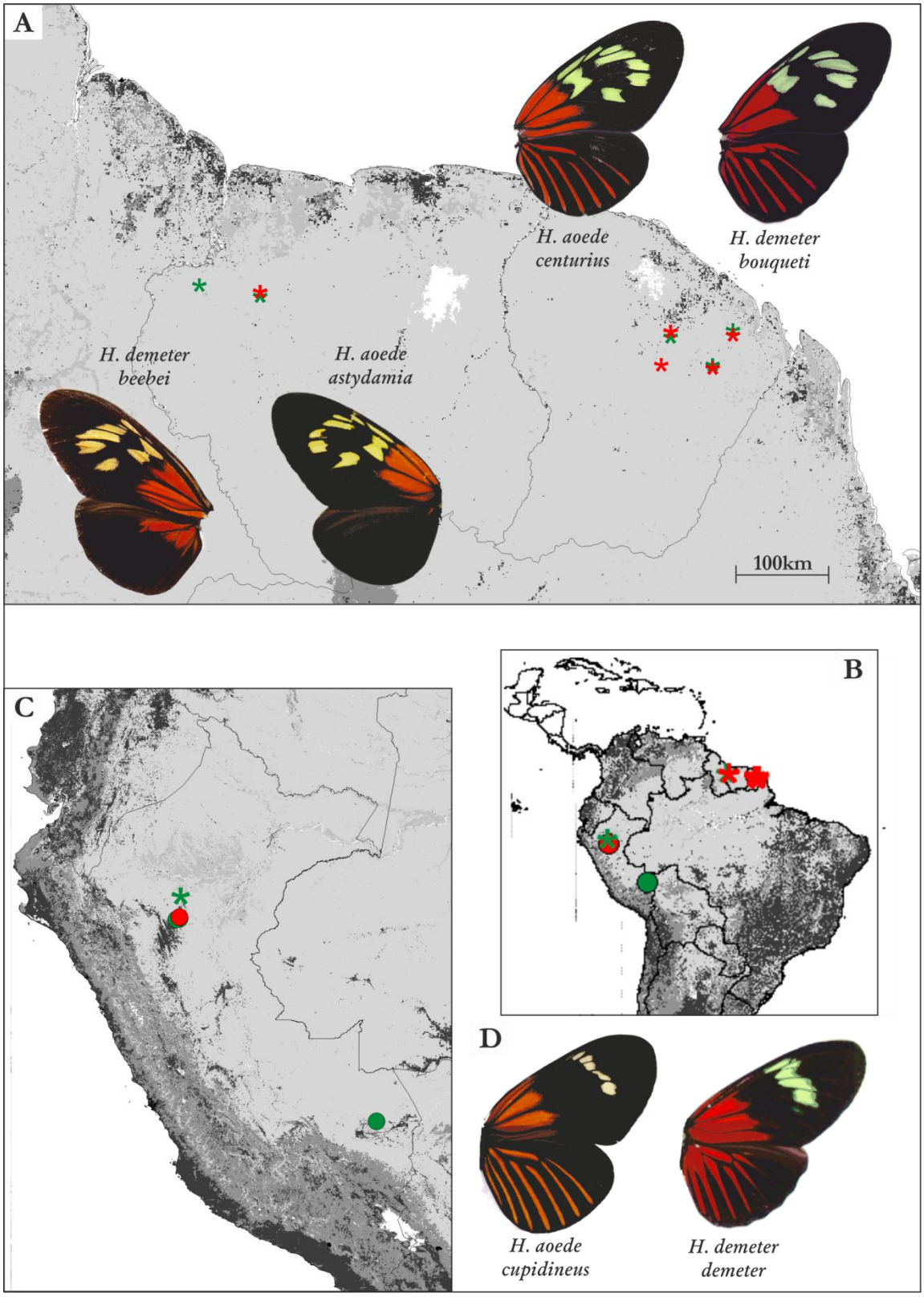
These colour pattern loci, *optix, cortex* and *WntA* are hotspots for the repeated convergent evolution of adaptive phenotypic variation across multiple *Heliconius* and non-*Heliconius* species. In addition, variation within species means that these colour pattern evolution hotspots can be identified as clear islands of divergence across hybrid zones, this has been seen time and again in population genomic studies in both *H. erato* and *H. melpomene* (Baxter *et al.* 2010; Counterman *et al.* 2010; Nadeau *et al.* 2013, 2014). Across these intraspecific hybrid zones, strong selection prevents gene flow at these adaptive colour pattern loci, while the rest of the genome can flow freely (Wu 2001). At the

moment this work has mostly just focussed on a small number of taxa that are either in the *H. melpomene*/silvaniform or *H. erato* clades. The genetic control of colour pattern in other species outside of these two clades, but that still form mimicry rings with these two species, are in contrast, generally unknown. As the example of Mexican cave tetras shows it is quite feasible that other species of *Heliconius* may well have arrived at different evolutionary answers to control colour pattern that utilise a different genetic toolkit.

Across the Amazon basin, eleven species from the tribe Heliconiini form a dennis-rayed mimicry ring. These include both *H. erato* and *H. melpomene*, as well as the silvaniform species *H. elevatus* and the species *H. timareta*. However, it also includes *Heliconius* species that analyses have shown are more distantly related from these clades (Kozak *et al.* 2015): *H. demeter, H. eratosignis, H. aoede, H. xanthocles, H. egeria* and *H. burneyi,* as well as *Heliconius Eueides tales* In the Guianas, seven of the eleven dennis-rayed Amazonian species form roughly concordant intraspecific hybrid zones, with the red hindwing rays pattern element found in colour pattern races from French Guiana and eastern Suriname, but absent in colour pattern races western Suriname and Guyana (Brown *et al.* 1974). As previously mentioned, the control of this colour pattern change has previously been established in both *H. melpomene* and *H. erato*. However, this is not the case for the five other species, in which the genetics of colour pattern have never before been investigated. With genomes for both *H. melpomene* and *H. erato* now available, the resources are now there to explore the genetic control of colour pattern in other species, an excellent opportunity to begin understanding the repeatability of evolution across this genus more widely.

Figure 2.1 (next page) – Locations of sampling sites of *H. aoede* (green) and *H. demeter* (red) across South America, A) Locations in French Guiana and Suriname. Top wings from left to right, *H. aoede centurius* and *H. demeter bouqueti* with orange dennis and hindwing rays, wings bottom wings from left to right *H. demeter beebei* and *H. aoede astydamia* with orange dennis only. B) Sites across South America. C) Sites across Peru. D) Wings from (left to right)

H. aoede cupidineus **and** ***H. demeter demeter*** **from Peru. * samples used for whole genome resequencing, circles; samples used for PCR amplification.**



The two species, *H. demeter* and *H. aoede*, are both divergent from *H. melpomene* and *H. erato*, as well as to each other, and so together with *H. melpomene* and *H. erato* they provide a limited but wide taxonomic sampling of the whole genera (Figure 1.1, from section 1.3.1). *H. demeter* is found in a clade which contains both *H. sara* and *H. sapho* with ~6 million years species in lacking pollen feeding (Penz & Krenn 2000). This had previously led them tobeing put in a different genus called *Neruda*, basal to *Heliconius* but more closely related than *Eueides*. More recently they have been placed within the *Heliconius,* at the base of the non-pupal mating clade that includes *H. melpomene*, the silvaniforms, *H. wallacei* and *H. doris*, with ~10 million years of evolution separating *H. aoede* from *H. melpomene* and more than that from *H. erato* (Kozak *et al.* 2015). In this chapter I investigated the genetic basis of colour pattern variation across the Guianese hybrid zones in *H. demeter* and *H. aoede* using a population genomics approach.

2.2 METHODOLOGY

2.2.1 **Test dataset**

As no published reference genome is available for *H. demeter* and *H. aoede*, analyses for these two species were carried out using reference genomes published for other *Heliconius* species (*H. erato* for *H. demeter* and *H. melpomene* for *H. aoede*) and *de novo* assemblies. In order to verify this *de novo* assembly approach, I first used a test dataset of *H. melpomene* from a hybrid zone in Peru. This was to see if regions of divergence could be located using a *de novo* assembled reference from one of these individuals. Furthermore, different parameters and assemblers could be trialled with this dataset. These could then be compared against the results when using the published *H. melpomene* (version 1.1) genome as a reference so that optimal parameters can be identified. This test dataset consisted of four *H. melpomene aglaope* (ERS235655, ERS235656, ERS235657, ERS235658) and *H. melpomene amaryllis* (ERS235651, ERS235652, ERS235653,

ERS235654) ~30x coverage whole-genome shotgun sequenced 100bp paired-end Illumina libraries (available from ENA; http:// www.ebi.ac.uk/ena/). These two subspecies form an intraspecific hybrid zone in Peru in which two loci control the major colour pattern variation. The *BD* locus that controls rays, dennis and band phenotypes, and the locus *Yb* controlling yellow forewing band and yellow hindwing bar phenotypes.

2.2.2 *De novo* assembly

The individual with the best coverage of these eight samples was chosen for *de novo* genome assembly (ERS235657). Before assembly the raw data was cleaned using cutadapt (Martin 2011), this removed adaptor sequences from the reads, trimmed low-quality ends (-q 20) and was used to discard remaining reads if their length was less than 15bp long. The program ABySS 1.3.1 (Simpson *et al.* 2009) was then used for genome assembly. K-mer size, the length in base pairs that the aligner splits the reads into prior to alignment was trialled at three different values, K30, K40 and K50. A second assembler, Platanus (Kajitani *et al.* 2014) which is designed for highly heterozygous genomes by using a variable K value that increases in a step wise fashion was also used to produce an assembly. Initial K was set to 32, while the step size was set to 10, default values were used for all other parameters.

Perl scripts were used to remove contigs below a read size of 100bp in each assembly (Appendix 8), as these were shorter than the read length. Following this, reciprocal Blasts within each assembly were carried out in order to identify highly similar contigs to be removed from each assembly (Appendix 9; Appendix 10). These highly similar contigs are most likely haplotype variants resulting from heterozygosity in the reference sample. Two different similarity thresholds were used; a relaxed threshold of 80 percent overlap and percentage identity, as well as a more stringent threshold of 95 percent overlap and percentage identity. Downstream scripts were then used to remove the shortest of each pair of these highly similar contigs (Appendix 11; Appendix 12). This gave a total of nine

different ABySS assemblies, with unfiltered assemblies, 95% filtered assemblies and 80% filtered assemblies at each of K30, K40 and K50, plus one Platanus assembly.

2.2.3 **Finding fixed differences**

For each *de novo* assembly, as well as the published *H. melpomene* genome (Dasmahapatra *et al.* 2012), BWA -aln (Li & Durbin 2009) was used to map reads from all eight *H. melpomene aglaope* and *H. melpomene amaryllis* samples back to the reference. These BAM files were then sorted using SAMtools, so that duplicate reads could be marked and removed before merging, both of which were done using PicardTools-1.100 (http://broadinstitute.github.io/picard/), these merged BAM files were then indexed using SAMtools. The Genome analysis Toolkit (GATK) 2.7-2 (McKenna et al., 2010) was then used to carry out realignment around indels before SNP calling with the GATK UnifiedGenotyper (DePristo *et al.* 2011). A Perl script from K. Dasmahaptra was then used to filter the resultant VCF files, so that only those SNPs without missing data across all individuals were retained. SNPs were called as missing if SNP quality > 30, genotype quality > 30, mapping quality > 20 and coverage > 5 < 150. Fixed differences between *H. melpomene aglaope* and *H. melpomene amaryllis* samples were then identified using a python script (by Simon Martin). A SNP was defined as fixed if all calls in the first population are the same, while this base is not found in any individual in the second population; further all calls in this second population are the same. A Perl script was then used to roughly identify the position of these fixed differences in the genome by blasting the *de novo* contigs with fixed SNPs against the reference *H. melpomene* genome (Appendix 13). R scripts were then used to plot fixed differences in 10kb sliding windows across the genome, using the *H. melpomene* genome as a reference. Fixed differences were termed singletons if they were not within 50kb of another fixed difference. This 50kb threshold was chosen, as linkage disequilibrium between two SNPs returns to

background levels at this distance, in the reference *melpomene* genome (Dasmahapatra *et al.* 2012).

2.2.4 *De novo* assembly quality metrics

As well as comparing the total number of fixed differences and their distribution across the genome, to the analysis using reads mapped to the published *H. melpomene* reference, a number of other metrics were also taken to determine the quality of an assembly. First, ABySS-fac was used to determine a number of metrics such as the number of contigs, cumulative length of all contigs, as well as the N50 of each assembly. After read mapping and BAM merging, Samtools' Flagstat was used to assess the number of reads that had successfully mapped. Furthermore, an estimate of redundancy within the assembly was also calculated. This redundancy estimates overlap between contigs, and therefore the heterozygosity remaining in the assembly, which might cause poor read alignment. Perl scripts (Appendix 14) were used to blast all contigs from an assembly against the reference genome. A second script was then used to calculate the proportion of bases covered by either one or multiple contigs (Appendix 15). From these blasts, a coverage measure could also be calculated for each assembly, this was the percentage of bases in the reference *H. melpomene* genome covered by at least one contig.

2.2.5 *H.* aoede and *H. demeter* sample collection and sequencing

H. aoede centurius and *H. demeter bouqueti* samples were collected by Mathieu Joron in French Guiana in 2009. *H. aoede astydamia* and *H. demeter beebei* samples were collected from Suriname in 2014. These samples were from western Suriname as the Guianese dennis-rayed hybrid zone runs through the east of this country (see appendix 1 for further details of the samples and locations). RNA-free genomic DNA was extracted for four samples of each subspecies to a concentration of approximately 15ng/µl from thoracic tissue using a Qiagen DNeasy Blood and Tissue Kit following the standard

protocol provided by the manufacturer. Libraries were prepared (by K Dasmahapatra) using TruSeq DNA PCR-Free Library Preparation Kits, with an insert size of approximately 350bp. Libraries were sequenced to ~40x coverage on an Illumina HiSeq 2000 instrument at the FAS Center for Systems Biology.

2.2.6 *De novo* assembly analyses

ABySS 1.3.1 was used to build *de novo* assemblies to be used as references, for both *H. demeter* and *H. aoede*. These were built using those parameters found to have given optimal results for the *H. melpomene* test dataset, K40 and the relaxed filtering of 80 percent overlap and percentage identity score. Again the two samples that had the greatest idealised coverage for each species were used to build the species *de novo* reference assembly. Idealised coverage was calculated by multiplying the number of reads by the read length, and then dividing by the length of the *H. melpomene* genome. BWA was used to map all eight samples of each species to the respective reference assembly, following which BAM files were sorted, duplicate reads were removed, these BAMs were then merged, and GATKs UnifiedGenotyper was used to call SNPs. Fixed differences were then found using the same python script (from Simon Martin) as for the test dataset. The chromosomal positions of these fixed differences, were then located using in house Perl scripts to BLAST *de novo* contigs against v1.1 of the reference *H. melpomene* genome for *H. aoede*, and the *H. erato* genome for *H. demeter* (Appendix 13).

2.2.7 Reference genome analyses

In addition to *de novo* genome assembly, reads for both *H. aoede* and *H. demeter* were also mapped to the phylogenetically closest published reference genome, *H. melpomene* for *H. aoede* and *H. erato* for *H. demeter,* in order to find fixed differences. Mapping was done using Stampy 1.0.27, with a substitution rate of 0.06 for *H. demeter* to *H. erato* and 0.10 for *H. aoede* to *H. melpomene.* Again BAM files were sorted, duplicate reads were

removed, these BAMs were then merged, and GATKs UnifiedGenotyper was used to call SNPs. Fixed differences were then found using the same python script (from Simon Martin) as had been used for the test dataset fixed differences.

Another analysis, the same as that described above for both *H. demeter* and *H. aoede,* but including an additional already sequenced sample of that species from Peru, was carried out as an outgroup analysis. Fixed differences were looked for between two groups; a group composed of Surinamese dennis-only samples, and a second group composed of samples from Peru and French Guiana. If the same loci are involved in the genetic control of the rays in both Peru and the Guianas, then this analysis should remove some of the fixed differences found between Surinamese and French Guiana samples that are not due to colour pattern differences.

2.2.8 **Fixed differences across analyses**

In order to get a complete picture of fixed differences across analyses, by seeing if the fixed differences identified in each analyses are the same, a Perl script (Appendix 16) was used to identify and extract 20bps of flanking sequence around each fixed difference in both the *de novo* and reference genome analyses. These flanking sequences were also in the same script reverse complemented in order to account for differences in orientation between *de novo* contigs and the reference genomes. These flanking sequences were then compared across analyses, where two flanking sequences each from the different analyses closely matched, with equal to or over 90% sequence identity, these fixed differences were identified as being the same. This gave an overall set of fixed differences for each species, composed of fixed differences identified in both the *reference* and *de novo* analyses, as well as those found in just one of these.

2.2.9 **Permutation tests of significance**

In order to determine the probability of finding a cluster of fixed differences of a given size in the genome, and to assess the significance of clusters, I used a permutation method to simulate the distribution of fixed differences in 10kb windows across the genome. For the reference genome analyses, every base with complete coverage across all individuals was noted and at these sites a fixed difference could be placed. For the *de novo* analyses each position across all contigs were allocated a unique genome position, these were ordered as they were found in the *de novo* reference. Following this, again every base with complete coverage across all individuals was noted and again fixed difference could only be placed at these sites. This allowed us to accurately simulate the effect of missing data in each window in each analysis. Following this, the same number of fixed differences as found in the empirical data, were then placed randomly across the genome. Each permutation thus gave a single genome with fixed difference clusters of various sizes. In total 1million permutations were carried out for each analysis to calculate the probability of finding a cluster of X SNPs in a genome. This probability works as a measure of the significance of a cluster of X SNPs given the empirically found complete coverage, genome size, and number of fixed differences. A significance level of $P < 0.001$ was used as the cut off above which a cluster of X size was deemed significant. This P-value significance level has been used in a similar context in previous work looking at divergence across genomes in *Helianthus* sunflowers (Renaut *et al.* 2013). It is also likely to be conservative given that this simulation approach does not take into account genetic linkage between fixed differences.

2.2.10 **Short range PCR amplicon sequencing**

With eight samples of each *H. demeter* and *H. aoede*, fixed differences from the *de novo* and reference analyses of each species were only known to be fixed over 16 alleles (eight diploid individuals). Targeted short range PCR sequencing was therefore used to

investigate whether fixed differences remained fixed over a larger sample size. This was done for regions showing significant divergence based on permutation tests, except for the most significant clusters of fixed differences and fixed differences at the *BD* locus. For these regions long range PCR was used, however the results of this are not included here due to a delay in sequencing. Primer locations for each amplicon were identified using multiple sequence alignments around fixed differences from each analyses. Perl scripts were used to make alignments of the variant call data from the reference genome analyses (Appendix 17). In addition to the focal taxa, the alignments included both the reference genome sequence (added with Appendix 18), as well as variant calls from an outgroup species, *H. wallacei*. The latter was included in an effort to try to design primers that where possible, were in conserved sequence blocks across taxa. This should lead to better primer performance across subspecies in *H. demeter* or *H. aoede*. Following this the VCF calls file was then converted to a fasta file of the alignment using a script from K. Dasmahapatra. *De novo* contigs with fixed differences from the *H. demeter* or *H. aoede* references were also included in alignments. These were aligned using BLAST, implemented with Perl scripts and then manually adjusted by eye (Appendix 19).

Following the building of each multiple sequence alignment, where possible conserved sequence blocks flanking the region of interest were located. In cases where fixed differences had only been found on *de novo* contigs conserved sequence blocks were still sought. Occasionally this was not possible and in these cases sequence that did not show conservation across *H. wallacei* was used. Consensus sequences made from both the variant calls from the reference individual of the target species and *de novo* contigs were used as the PCR template in Primer-BLAST (Ye *et al.* 2012). Coordinates of target blocks were then input into Primer-BLAST, and primers that gave a maximum amplicon length of 800bp were searched for.

In total, five *H. aoede astydamia* from Suriname (including the four samples used in the whole genome analyses), 13 *H. aoede centurius* from French Guiana (including the four samples used in the whole genome analyses) and five *H. aoede cupidineus* from Peru (Figure 1.1) were used for amplicon sequencing (see appendices 3 and 4 for sample details). A total of five *H. demeter beebei* from Suriname (including the four samples used in the whole genome analyses), fifteen *H. demeter bouqueti* samples (including the four samples used in the whole genome analyses) and one *H. demeter ucayalensis* from Peru (Figure 1.1) were used for amplicon sequencing. RNA-free genomic DNA was extracted to a concentration of approximately 15ng/µl from thoracic tissue, using a Qiagen DNeasy Blood and Tissue Kit following the standard protocol provided by the manufacturer.

PCR amplifications were performed using 10µl reaction volumes: generally consisting of 5.7µl of autoclaved aquapure H_2O; 2µl of 5X Green GoTaq® Flexi Buffer (Promega); 0.6µl of 25mM $MgCl_2$,; 0.2 µl of 10µM DNTPs, 0.2µl of each 10µm primer, 0.1µl of GoTaq® G2 Flexi DNA Polymerase (Promega); and 1ul of genomic DNA. The standard PCR program consisted of an initial two minute denaturation at 95°C, followed by 35 cycles of three-steps: another 95°C denaturation for 45 seconds; a 45 second annealing step for which temperatures varied for different primer pairs (see appendix 2); and a 45 second extension step at 72°C, before a final five minute extension again at 72°C. For one amplicon, blanket annealing temperatures did not work well across all samples, some were run with the standard PCR program detailed above, while others were run using a touchdown program. This consisted of ten cycles starting with annealing temperatures at 70°C and going down 1°C each cycle, until 60°C where a further 25 cycles were run. Following PCR, amplicon products were visually checked using by running 1µl on 1% agarose gels. PCR products were then cleaned using microclean (made in-house), then cycle sequenced using standard protocols for the BigDye® Direct Cycle Sequencing Kit (Applied Biosystems, UK), before finally being sequenced on an ABI 3730 sequencer (Applied Biosystems, UK).

Chromatograms of sequence data were checked and edited using SeqTrace 0.9.0 (Stucky 2012) and aligned in the program BioEdit (Hall 1999), using ClustalW (Thompson *et al.* 1994). Genotype calls for the relevant *de novo* contig were then extracted from the VCF file, and the script from Appendix 17 was used to add sites with no data with IUPAC code N, this file was then converted to fasta format using a script from K. Dasmahapatra. The PCR alignment was then added to this VCF based alignment and secondarily checked by visual inspection. The genotype information could then be checked at each fixed difference across all individuals that had been successfully sequenced. Where there had been incomplete coverage in the *de novo* analysis, and therefore missing fixed differences, calls from the whole genome analysis were used for those samples to check if differences were still fixed. Likewise for the *H. aoede* amplicons, where a call for a fixed difference was available from the whole genome sequenced *H. aoede cupidineus* from Peru (that had been mapped to the *H. melpomene* reference) this information was also utilised.

2.2.11 Characterising regions of divergence

Following the discovery of regions of clustered fixed differences across the genomes of both *H. demeter* and *H. aoede,* these regions were investigated in LepBase (http://ensembl.lepbase.org/index.html) using the *H. melpomene* or *H. erato* reference genomes. This was to identify whether these regions contained any annotated genes. Following the discovery of a gene, the nucleotide sequence of this was copied from the blast function in Lepbase into that of Flybase (http://flybase.org/). Tblastx was then used to determine whether any known function was ascribed to the ortholog in the well characterised *Drosphila melanogaster* geneome, in order to identify if it might be a good candidate for a role in colour pattern control. A small number of genes from the *H. erato* genome were also compared to their orthologs in the *H. melpomene* genome. This was also done using tblastx and was done through the blast function in LepBase.

2.3 RESULTS

2.3.1 *H. melpomene* test dataset analyses

A total of 2219 fixed differences found when mapping all four *H. melpomene aglaope* and *H. melpomene amaryllis* samples to the published *H. melpomene* reference genome, with 94% of these found on just three of the 4309 scaffolds (Figure 2.2). Two of these three scaffolds, HE670865 which had 53% of all fixed differences, and HE667780 which had 34% of all fixed differences, contained the red and yellow colour pattern controlling loci respectively. A third scaffold HE671488 had a further 7% of the fixed differences, this scaffold has previously been found to be divergent between *H. melpomene* colour pattern races and is thought to perhaps be associated with altitude (Nadeau *et al.* 2014).

The results for the *de novo* genome quality metrics for each analysis with each assembly are shown in Table 2.1. In the *de novo* assembly analyses, the proportion of fixed differences found on each of these three scaffolds were similar to the proportions found using the *H. melpomene* reference genome, with approximately 90% of all fixed differences found on the three scaffolds; HE670865, HE667780, HE671488 (Fig 3.1). However, the total number of fixed differences found across the genome was less than half the number found using the *H. melpomene* reference genome, this is likely explained by missing parts of the genome in the *de novo* assembly, as well as increased heterozygosity which leads to poor mapping. Of the three K values tested, K40 consistently found the most fixed differences across any given redundancy filtering level. However, N50 increased with K, so that K50 gave the highest N50. Redundancy also increased with K, but filtering out redundant contigs across all values of K consistently increased the total number of fixed differences, as well as the N50. The assembly built using Platanus showed low redundancy, but had a low N50 and did not recover as many fixed differences as some of the filtered ABySS

K	Filtering	N50	% Coverage	% Redundancy	% Mapped reads	Total fixed diffs.	% fix diffs on HE671488	% fix diffs on HE667780	% fix diffs on HE670865	Deviation from ref %
H. melpomene reference v1.1	-	196221	100	0	-	2219	6.67	33.80	53.27	-
K50	unfiltered	1912	84	61	69.85	807	7.56	28.13	53.90	7.20
	95	1849	83	40	69.34	910	7.47	27.69	55.05	8.70
	80	2202	80.4	20	64.18	948	7.28	29.01	54.01	6.14
K40	unfiltered	1544	82	49.3	60.33	905	8.95	27.85	52.71	8.79
	95	1490	81.3	34.6	60.14	964	8.40	28.53	52.70	7.57
	80	**1751**	**79**	**17.6**	**57.84**	**1004**	**8.17**	**30.08**	**51.49**	**6.99**
K30	unfiltered	955	77	27.8	51.55	885	7.68	29.94	53.11	5.03
	95	930	77	23	51.57	902	7.54	29.71	52.22	6.01
	80	1047	75	11	50.62	909	7.59	30.14	52.04	5.81
Platanus	-	1113	76	18.93	50.47	921	8.58	28.12	54.18	6.67

Table 2.1 - The results of the *de novo* genome quality metrics for each analysis with each assembly, in comparison to the analysis carried out using the *H. melpomene* reference genome. In bold are the results for the analysis using a *de novo* reference built with assembly parameters that were chosen to be used for the *H. demeter* and *H. aoede* analyses. The deviation from reference percentage was calculated by summing the differences in the percentage of fixed difference on scaffolds HE671488, HE667780 and HE670865 in the analysis with the *H. melpomene* reference, from the percentage of fixed differences on these scaffolds from the analysis using the respective *de novo* assembly.

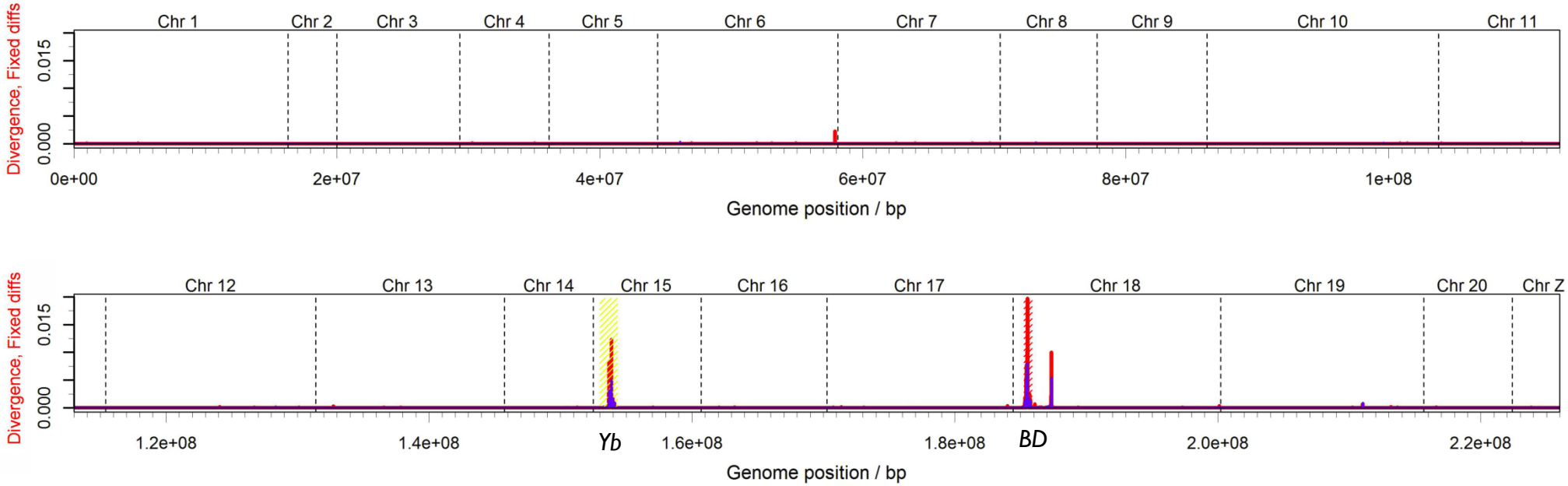


Figure 2.2 - Divergence between *H. melpomene aglaope* and *H. melpomene amaryllis* across the genome, shown as the proportion of fixed differences fixed in a 10kb (5kb sliding) window. Red shows divergence in analysis using published reference *H. melpomene* genome, blue shows divergence from the analysis using the K40 *de novo* assembly filtered with 80 percent overlap and percentage identity. Yellow and red shading show *Yb* and *BD* colour pattern scaffolds respectively.

assemblies. Filtering was generally successful reducing the percentage redundancy score, while having minimal effect on the coverage score, indicating that as hoped filtering targeted heterozygous regions. Overall the ABySS assembly with a K value of 40, and with post-hoc filtering of redundant reads with percentage overlaps and percentage identity scores over 80 was determined to be best. This assembly parameter set was chosen because i) the analysis conducted with this assembly found the most fixed differences, while the proportions of these fixed differences on the three scaffolds were similar to those found when using the *H. melpomene* reference genome, ii) this assembly had a reasonably good N50, and iii) the assembly had a relatively low redundancy score.

2.3.2 **Divergence across the *Heliconius aoede* genome**

The *H. aoede* sample with the highest idealised coverage was MJ09-4015 with ~104x coverage; this sample was therefore used to build the *de novo* reference assembly. This final assembly contained 381,498 contigs, had a genome size of ~259.9mb, and an N50 of 1953bp. In total, 263 fixed differences were found between the rayed *H. aoede centurius* from French Guiana and the non-rayed *H. aoede astydamia* from Suriname when using this reference. Fifteen fixed differences were located on scaffolds unmapped in the *H. melpomene* genome, of the other 248, 101 were singletons, defined as not being within 50kb of another fixed difference. The other 147 fixed differences were in groups of two or more. When reads were aligned directly to the reference *H. melpomene* genome 67 fixed differences were identified. Again a similar proportion of these fixed differences, thirty-two, were singletons. Permutation tests simulating the data mapped to the *H. melpomene* reference showed that only clusters of three fixed differences or more had a $P < 0.001$ (singletons had $P = 1.000$, clusters of two fixed differences $P = 0.106$).

For the *de novo* genome mapped analysis permutation tests found that only clusters of four differences or more had a $P < 0.001$ (singletons had $P = 1.000$, clusters of two fixed differences $P = 0.706$, three $P = 0.004$). Thirty-five of the fixed differences were identified

in both analyses, giving an overall total of 295 fixed differences across the genome (Figure 2.3). One cluster of these fixed differences was found on chromosome 18, specifically on the *BD* scaffold (HE670865), close to the location of the *rays* locus in *H. melpomene*. Association mapping has located this locus to between 333kb and 372kb along HE670865 (Wallbank *et al.* 2016). However, the largest cluster of fixed differences was located on chromosome 10 (scaffold HE670875), composed of 62 fixed differences within a wider ~50kb region, with 54 of these within a narrower ~20kb region. The second largest cluster of fixed differences was located on chromosome 8 (HE671576) composed of 17 fixed differences in a ~30kb region. Additional significant clusters of fixed differences are detailed in Table 2.2.

Chrom.	Scaffold	Position	Both analyses	*De novo*	*H. melpomene* reference	*H. melpomene* with outgroup
1	HE671150	44783 - 49763	4	4*	0	0
8	HE671576	100468 - 128504	17	17*	3*	3
10	HE670875	24652 - 79066	62	56*	15*	11
12	HE672075	786327 - 787042	5	4*	3*	2
16	HE671862	148053 - 149466	7	7*	0	0
18	HE670865	357138 – 393357	8	7*	5*	4
19	HE670348	68648-76602	7	7*	0	0
Z	HE671266	168,619-182405	9	9*	3*	3

Table 2.2 – Significant clusters of fixed differences found between rayed *H. aoede astydamia* and non-rayed *H. aoede centurius* in each analysis and overall. Locations are shown in reference to the *H. melpomene* genome v1.1. * indicates fixed difference cluster was found to be significant in permutation test for that analysis.

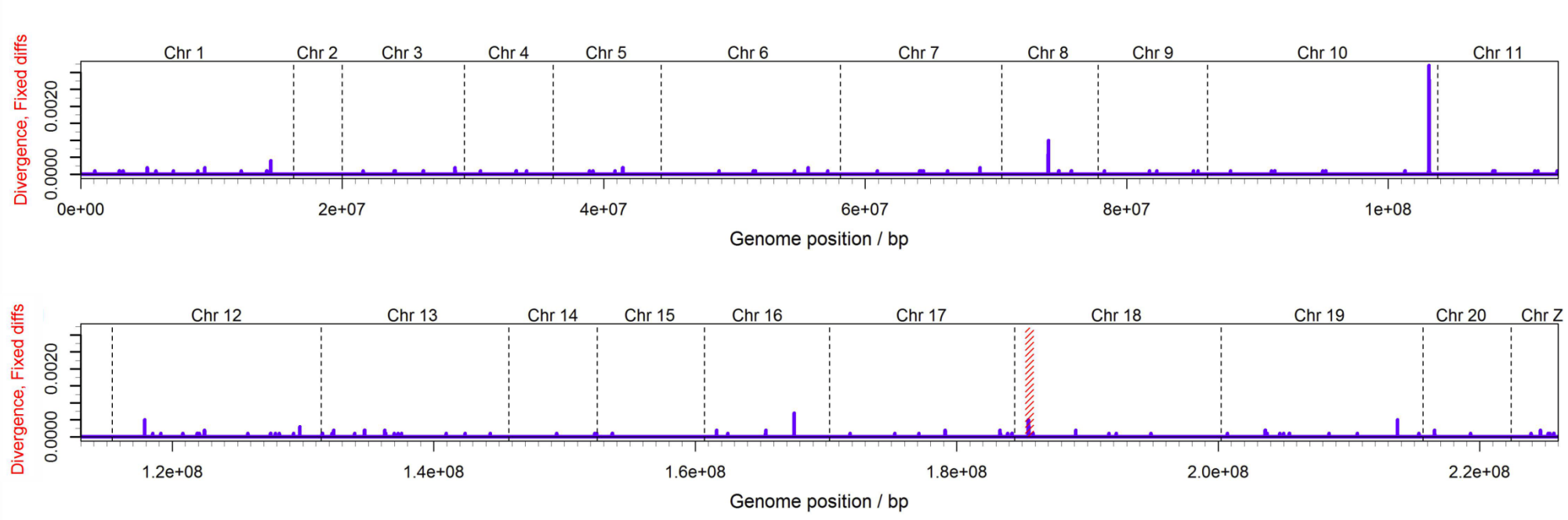


Figure 2.3 – Divergence across the genome as the frequency of fixed differences per 10kb window (with 5kb slide), between rayed *H. aoede centurius* versus non-rayed *H. aoede astydamia* oriented using the *H. melpomene* reference genome. Plot includes fixed differences from both the analyses using published *H. melpomene* reference genome and the analyses with *de novo* reference. Red shaded rectangle indicates BD locus.

When a Peruvian outgroup individual was included in the *H. melpomene* reference analysis, so that only SNPs fixed between rayed *H. aoede centurius* and *H. aoede cupidineus* from French Guiana and Peru, and non-rayed *H. aoede astydamia* from Suriname were counted, a large number of fixed differences dropped out leaving only 37 fixed differences. These included eleven on scaffold Chromosome 10 (scaffold HE6710875) and four in the *BD* region (scaffold HE670865), as well as clusters of three fixed differences on HE671266 and HE671576 and two on HE672075. These clusters were therefore selected for short range PCR sequencing to check that these SNPs were really fixed across a larger sample size.

2.3.3 Divergence across the *Heliconius demeter* genome

Of the eight *H. demeter* samples, the sample with the highest idealised coverage was 2014-59 with ~100x coverage; this sample was therefore used to build the *de novo* reference assembly. This final assembly contained 475645 contigs, had a genome size of ~307.4mb, and an N50 of 2145bp. In total, 271 fixed differences were found between rayed *H. demeter bouqueti* from French Guiana and non-rayed *H. demeter beebei* from Suriname when using this reference. Of these fixed differences, 148 were singletons, and so likely to be result of small samples sizes and drift. When aligning reads to the *H. erato* reference genome 190 fixed differences were identified. This was almost three times the number of fixed differences found when mapping *H. aoede* reads to the *H. melpomene* genome. This likely reflects the closer phylogenetic relationship between *H. demeter* and *H. erato.* Of these 190 fixed differences, 92 were singletons. Permutation tests using simulating the data mapped to the *H. erato* reference found that only clusters of four fixed differences or more had a $P < 0.001$ (singletons had $P = 1.000$, clusters of two fixed differences $P = 0.889$, three $P = 0.018$). This was true for the *de novo* genome mapped analysis as well, with only clusters of four differences or having $P < 0.001$ (singletons had $P = 1.000$, clusters of two fixed differences $P = 0.650$, three $P = 0.003$).

85 of these fixed differences were identified across both analyses, giving an overall total of 376 fixed differences across the genome (Figure 2.4). Just as in the *H. aoede* analysis, a cluster of fixed differences was found on Chromosome 18 in the locus known to control red patterning, the D locus in *H. erato* (on scaffold H*erato*801). This was composed of 10 fixed differences. However, the largest cluster of fixed differences was again found outside of this region, on Chromosome 2 (scaffold H*erato*0206) composed of 51 fixed differences in a ~50kb region. The majority of these were even more focussed, into two clusters. Thirty-five fixed differences in a ~1kb region (488371bp - 489455bp on H*erato*0206) and a second cluster of fifteen fixed differences in another ~1kb region (494562 - 495988bp H*erato*0206). A number of other smaller significant clusters of fixed differences were also found across a number of other chromosomes (see Table 2.3).

Chrom.	Scaffold	Position	Overall analysis	*De novo*	*H. erato* reference	*H. erato* with outgroup
1	0101	2726827 - 2727749	5	5*	5*	1
2	0206	447016 - 495988	51	26*	37*	3
6	0606	1177591 - 1178909	5	0	5*	1
8	0801	2784851 - 2784944	4	4*	3	3
9	0901	7845425 - 7845573	5	5*	1	1
13	1301	8885524 - 8886129	4	4*	1	0
16	1601	298961 - 299097	4	0	4*	0
18	1801	1350386 - 1351447	10	5*	6*	1
19	1910	2149913 - 2161999	5	5*	0	0
21	2101	11908509 - 11942857	5	1	5*	0

Table 2.3 – Significant clusters of fixed differences found between rayed *H. demeter bouqueti* and non-rayed *H. demeter beebei*, in each analysis and overall. Locations are shown in reference to the *H. erato* genome v1. * indicates fixed difference cluster was found to be significant in permutation test for that analysis.

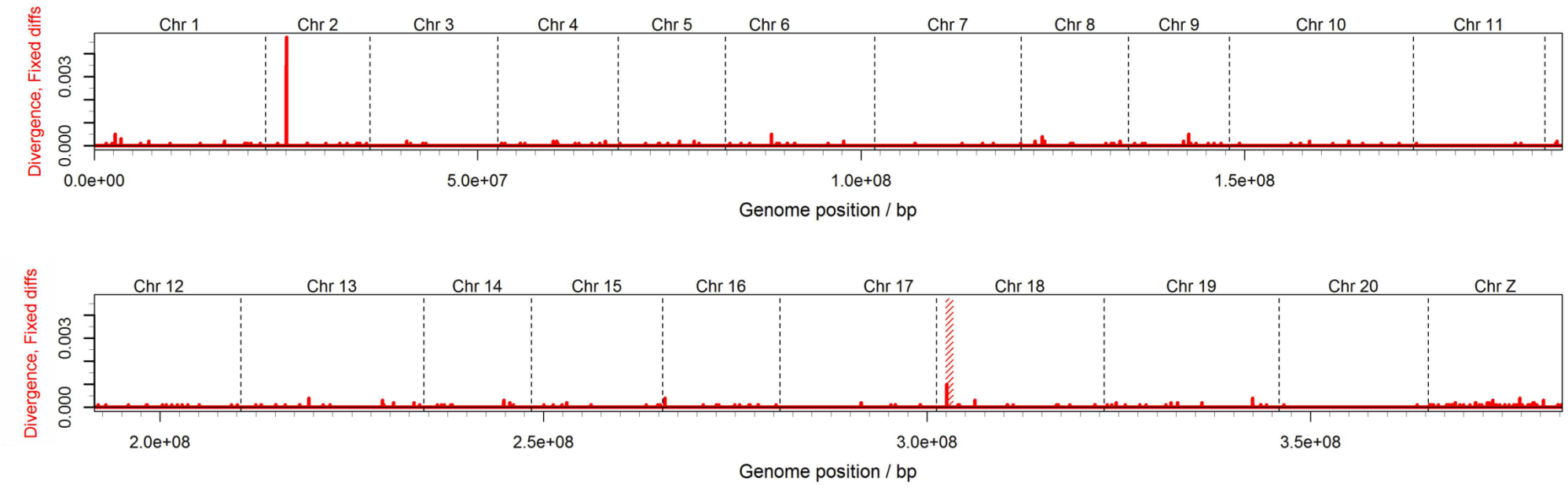


Figure 2.4 – Divergence across the genome as the frequency of fixed differences per 10kb window (with 5kb slide), between rayed *H. demeter bouqueti* versus non-rayed *H. demeter beebei* oriented sing the published *H. erato* reference genome. Plot includes fixed differences from both the analyses using published *H. erato* reference genome and the analyses with *de novo* reference. Red shaded rectangle indicate BD locus.

When a Peruvian outgroup individual was included in the *H. erato* reference analysis, so that only SNPs fixed between rayed *H. demeter bouqueti* and *H. demeter demeter* from French Guiana and Peru, and non-rayed *H. demeter beebei* from Suriname were counted, a large number of fixed differences dropped out leaving only 61 fixed differences. Of these, only one fixed difference from the cluster on scaffold 1801 remained, while just three remained from the largest cluster on scaffold 0206, another three were still found on scaffold 0801. All clusters with fixed differences from this outgroup analysis were among those selected for short range PCR sequencing, to check that these SNPs were really fixed across a larger sample size.

2.3.4 **Short range PCR amplicon sequencing**

Regions showing the greatest divergence along with fixed differences around the *BD* locus were sequenced using long range PCR, however the results of this are not included here due to a delay in sequencing. Short range PCR was though used to sequence other regions with fixed differences to check that these SNPs were really fixed using more samples. Three amplicons were successful sequenced for three of the *H. aoede* clusters of fixed differences. Of these fixed differences, only those on one amplicon remained fixed with an extended sample size. This set of fixed differences was found on the *H. melpomene* scaffold HE671576 and *de novo* contig 3735109. These fixed differences together were part of the second largest cluster of fixed differences across the genome (Table 2.2). Amplicon sequencing showed that even across a larger sample size these differences remained fixed (Table 2.4).

The other two amplicons covered two SNPs on the *H. melpomene* scaffold HE671266 (*de novo* contig 347519), and five fixed differences from *H. melpomene* scaffold HE672075, that had been found to be fixed from the whole genome analyses without an outgroup. All of these were found not to be fixed over a larger sample size, this was mainly due to one extra sample of *H. aoede astydamia* which was homozygous for the French Guiana allele at

all fixed differences across these two loci. In addition, at two fixed differences, one from either amplicon, some *H. aoede centurius* samples were also found to have Surinamese alleles. Fixed differences in these two clusters were therefore not found to be fixed over larger sample sizes (see Table 2.4), suggesting they likely do not play a role in the control of colour pattern.

H. melpomene Scaffold	Scaffold Position	Contig	Sample size (without | with outgroup)	Percent fixed across (without Peru Outgroup)	Percent fixed across (with Peru Outgroup)
HE671576	101154	3735109	17 | 22	100	100
HE671576	100468	3735109	17 | 21	100	100
HE671576	101244	3735109	17 | 22	100	100
HE671576	100613	3735109	17 | 21	100	100
HE671576	100643	3735109	17 | 21	100	100
HE671266	176803	3479519	16 | 22	93.75	95.45
HE671266	176727	3479519	16 | 22	75	68.18
HE672075	786962	3722063	16 | 22	93.75	95.45
HE672075	786608	3722063	16 | 21	93.75	90
HE672075	786327	3722063	16 | 22	93.75	95.45
HE672075	787024	3722063	16 | 21	68.75	70
HE672075	787042	NA	16 | 21	93.75	95

Table 2.4 – SNPs found to be fixed across the genome in *H. aoede* WG analyses, with expanded sample sizes from targeted PCR amplicon sequencing, showing location, new sample size and across what percent of individuals the SNP remained fixed. White, fixed differences unique to the *de novo* analysis; light grey, unique to the *H. melpomene* reference genome analysis; dark grey, found in both analyses.

Eight primer pairs were used to successfully sequence amplicons containing six clusters of fixed differences in *H. demeter*. Across the total of 25 fixed differences checked in *H. demeter* none remained fixed with the expanded sample size, often without needing to expand sample size by a very large amount (Table 2.5), suggesting they do not play a role in the control of colour pattern. The first two of these primer pairs covered all five fixed

differences on scaffold H*erato*0101. Amplicon sequencing successfully expanded samples sizes for both of these primer pairs, with all of these *H. demeter bouqueti* samples from French Guiana found to have the allele previously only found in the Surinamese samples. (see Table 2.5). The sample size was also successfully expanded for four of five fixed differences on the scaffold H*erato*0606. Two of these were found to have the allele previously only found in the *H. demeter beebei* samples from Surinamese. The next two primer pairs covered three of four fixed differences on the scaffold H*erato*0801. This time the majority of these new French Guiana samples were found to have the allele previously only found in the *H. demeter beebei* samples from Suriname.

Amplicon sequencing successfully expanded samples sizes for all five fixed differences found on the scaffold H*erato*0901. The additional Peruvian sample sequenced was found to have the same genotype across all five SNPs as the original four rayed samples from French Guiana used in the whole genome analysis. However, the additional French Guiana samples were found to have the allele previously only found in the *H. demeter beebei* samples from Suriname. Amplicon sequencing also successfully expanded samples sizes for all four fixed differences found on the scaffold H*erato*1601. One of the additional samples from French Guiana was found to have the allele previously only found in the *H. demeter beebei* samples from Suriname, and so these SNPs were not fixed over a larger sample size. The final amplicon covered four of five fixed differences on the scaffold H*erato*2101. Again these SNPs were not fixed over a larger sample size

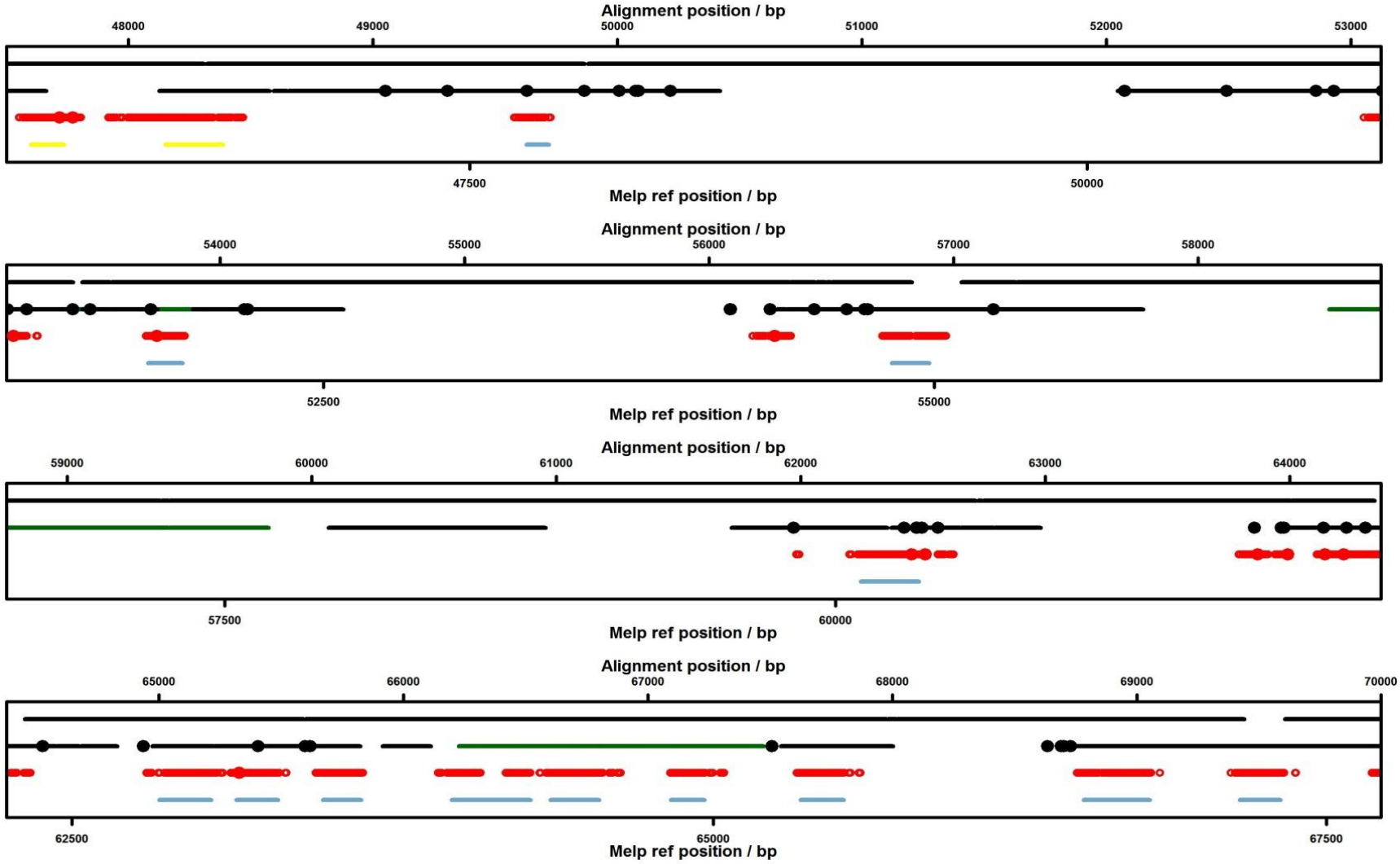
Table 2.5 (next page) - SNPs found to be fixed across the genome in *H. demeter* WG analyses, with expanded sample sizes from targeted PCR amplicon sequencing, showing location, new sample size and across what percent of individuals the SNP remained fixed. (*coverage was not complete across samples in the *de novo* genome population genomics analysis, so SNP had not been found to be fixed in *de novo* analysis) White, fixed differences unique to the *de novo* analysis; light grey, unique to the *H. melpomene* reference genome analysis; dark grey, found in both analyses.

H. erato Scaffold	Contig	Scaffold Position	Sample size (without | with outgroup)	Percent fixed across (without Peru Outgroup)	Percent fixed across (with Peru Outgroup)
H*erato*0101	5948228	2725971	11	72.72	NA
H*erato*0101	5948228	2726241	10	80	NA
H*erato*0101	5948228	2727523	16 | 17	75	76.47
H*erato*0101	5948228	2727640	16 | 17	68.75	70.59
H*erato*0101	5948228	2727759	14 | 15	78.57	80
H*erato*0606	1536*	1177591	11	81.81	NA
H*erato*0606	1546*	1177601	11	81.81	NA
H*erato*0606	1554*	1177609	11	81.81	NA
H*erato*0606	1565*	1177620	11	81.81	NA
H*erato*0801	6104138	2784073	16 | 17	62.5	64.70
H*erato*0801	6104138	2784107	16 | 17	62.5	64.70
H*erato*0801	6104138	2784114	16 | 17	56.25	58.82
H*erato*0901	5958220	7845425	10 | 11	90	90.90
H*erato*0901	5958220	7845496	11 | 12	81.81	83.33
H*erato*0901	5958220	7845501	11 | 12	81.81	83.33
H*erato*0901	5958220	7845555	11 | 12	81.81	83.33
H*erato*0901	5958220	7845573	11 | 12	90.90	91.66
H*erato*1601	397*	298961	12	91.66	NA
H*erato*1601	404*	298968	12	91.66	NA
H*erato*1601	407*	298971	12	91.66	NA
H*erato*1601	535*	299097	12	91.66	NA
H*erato*2101	2597*	11942634	17 | 18	64.71	61.11
H*erato*2101	2581*	11942650	17 | 18	58.82	55.55
H*erato*2101	2532*	11942699	17 | 18	58.82	55.55
H*erato*2101	2372*	11942857	17 | 18	64.71	61.11

2.3.5 **Characterising regions of divergence**

Genes around or containing significant clusters of fixed differences were investigated by looking at orthologs within the *Drosophila melanogaster* genome. For *H. aoede* the largest cluster of fixed differences was on chromosome 10 on scaffold HE670875, of these 51 of the 62 found, were within the gene *cardinal*. This gene has functions described as both Heme binding and in peroxidase activity. The gene is also described as having a role in number of biological processes, namely kynurenine metabolic process, negative regulation of gene silencing by RNA, positive regulation of neuron death and in the ommochrome biosynthetic process (http://flybase.org/reports/FBgn0263986.html). This last process makes *cardinal* a good candidate for a role in colour pattern control, as these ommochrome pigments used in insect eyes are also those used in *Heliconius* wing patterning (Reed & Nagy 2005; Ferguson *et al.* 2011). Interestingly The majority of these fixed differences were concentrated within introns. While many of these are unlikely to be functional, but caused by hitchhiking with those which are, the lack of divergence in protein coding sequence suggests that this is conserved across colour pattern forms. It seems therefore, that if *cardinal* is involved in the non-rays phenotype in *H. aoede* that this has evolved through changes in *cis*-regulatory modifiers, rather than protein coding changes.

Figure 2.5 (next page) – Alignment of *cardinal* contigs against the *H. melpomene* genome v1.1 (Scaffold HE670875). Top track: *H. melpomene* scaffold, gaps indicate insertions in *H. aoede* contigs. Second track: positions of *de novo* contigs, black have fixed differences (black circles), green lack fixed differences. Third track: sites with complete coverage (red) from reads aligned to *H. melpomene* genome, red circles indicate fixed differences. Bottom track: *cardinal* exons shown in blue, exons from neighbouring genes shown in yellow.


Alignment position / bp
48000
49000
50000
51000
52000
53000
47500
50000
Melp ref position / bp
Alignment position / bp
54000
55000
56000
57000
58000
52500
55000
Melp ref position / bp
Alignment position / bp
59000
60000
61000
62000
63000
64000
57500
60000
Melp ref position / bp
Alignment position / bp
65000
66000
67000
68000
69000
70000
62500
65000
67500
Melp ref position / bp

The other main cluster of fixed differences was found on chromosome 8, on the scaffold HE671576. Two genes appeared to contain most of these 17 fixed differences on the *H. melpomene* genome HMEL016481 and HMEL016483, in addition one fixed difference was found within HMEL016482. These three *H. melpomene* genes hit two targets within the *D. melanogaster* genome, suggesting that HMEL016481 and HMEL016482 are paralogous, or have been incorrectly defined during genome annotation, and are actually part of one larger gene. This gene that both hit, *Easter*, has a function in Serine peptidase activity and has a function in dorsal/ventral axis specification and zymogen activation. In addition, the gene that HMEL016483 hits is Spatzle-Processing Enzyme (*SPE*) again a gene described as being involved in serine-type endopeptidase activity and appears to have roles in various forms of immune defence response.

The *BD* scaffold HE670865 on Chromosome 18, is known to contain the locus involved in rays control in *H. melpomene*. This whole region is a gene desert, with no genes found between 311kb and 438kb, so none of these genes were checked in *D. melanogaster*. However, upstream of these fixed differences at 438kb on scaffold HE670865 is the transcription factor *optix,* this is known to be involved in rays patterning in other *Heliconius* species (Reed *et al.* 2011; Martin *et al.* 2014b). Five of the eight fixed differences on this scaffold were at ~390kb which is just up-stream of a region putatively described as a regulator enhancer (~330kb-370kb) that controls the rays phenotype in *H. melpomene*, while the other three fixed difference were in fact found within this *rays* region, this proximity suggests that the region of divergence on the BD locus in *H. aoede*, is broadly homologous to the *rays* module in *H. melpomene*.

Other clusters of fixed differences were smaller and have been found through PCR not to remain fixed between colour patterns groups when sample size is increased (Table 2.4). However these genes did still show elevated divergence across the hybrid zone and so may still be adaptive. Genes close to these clusters are described in (Table 2.6).

Scaffold	Gene Position	*H. melpomene* name	*Drosophila* name	TblastX e-value	Function
HE671150	24692 - 48655	HMEL010910	CG9541	8.43494e-89	ATP binding; adenylate kinase
HE671576	92,924-95,799	HMEL016480	CG12948	7.99031e-06	No functional information
HE671576	96,100-101,290	HMEL016481	*Easter* (CG4920)	9.11673e-35	Serine peptidase activity
HE671576	105,241-109,375	HMEL016482	*Easter* (CG4920)	2.36589e-22	Serine peptidase activity
HE671576	112,748-129,481	HMEL016483	*Spatzle-Processing Enzyme* (CG16705)	3.3197e-19	Serine-type endopeptidase activity
HE671576	130,224-133,368	HMEL016484	CG18109 (Not good match)	1.0867	Gamma tublin creation
HE670875	29,956-40,310	HMEL009167	*SpellChecker1*	0.000699592	ATP binding
HE670875	40,612-46,570	HMEL009168	CG1749	1.70861e-79	Mo-molybdopterin cofactor sulfurase activity
HE670875	47,734-68,205	HMEL009169	*cardinal* (CG6969)	3.42408e-88	Heme binding; peroxidase activity
HE670875	82,953-85,279	HMEL009170	CG5001	8.33286e-125	Unfolded protein binding
HE672075	37,150-42,762	HMEL016801	No gene	0.000142675	-
HE672075	139,410-140,015	HMEL016802	No gene	0.269389	-
HE671862	132,299-139,485	HMEL008238	CG43867	1.15447e-10	No functional information
HE670348	55,957-88,526	HMEL006026	*ninaB* (CG9347)	2.43216e-43	Carotenoid dioxygenase activity; retinal isomerase activity
HE671266	153,622-198,729	HMEL012199	SERCA (CG3725)	2.58337e-60	Calcium-transporting ATPase activity; metal ion binding; nucleotide binding

Table 2.6 – Location and functional information (from *D. melanogaster*) of genes at or near clusters of fixed differences found between *H. aoede astydamia* and *H. aoede centurius*.

Again in *H. demeter* the fixed differences found on chromosome 18, scaffold H*erato*1801 were found far from any genes, with the closest being H*erato*1801.64 ~(99kb away, at 1,239,943 - 1,251,211 on scaffold H*erato*1801) and H*erato*1801.65 (~77kb away, at 1,427,434 - 1,435,218). These genes did not come up with any clear hits against *D. melanogaster*, but hit the genes *optix* (Hmel001028; 438,423 - 439,107 on HE670865) and HMEL001014 (306,696 - 311,266 on HE670865) respectively, in the *H. melpomene* v.1.1 genome. This again places these fixed differences in and around the gene *optix*, and more specifically within the *rays* locus known from *H. melpomene,* and just 30kb away from a putative *rays* module in *H. erato* (Van Belleghem *et al.* 2016).

The largest peak of fixed differences in *H. demeter* was, in contrast, found within a region of the *H. erato* genome that had a number of genes in and around it. However, when blasted to the *D. melanogaster* genome these genes appeared to have no clear orthologs, with one appearing to hit a repeat region. Therefore the suitability of these genes as candidates was unclear. A peak of divergence has though been found between postman and dennis-rayed races of *H. erato* (Nadeau *et al.* 2014; Van Belleghem *et al.* 2016) that covers a wide region including the peak of divergence seen in *H. demeter*. It is thought that this region may be an ancient inversion between the postman and rayed races. One possible explanation for the elevated divergence in *H. demeter* is thus that this is also the site of an inversion. However, at the moment this is hard to know, and alternatively it may be a repeat region, or in fact be under colour pattern selection. Interestingly, as in the *H. aoede* analysis a number of the genes in and around the smaller regions of clustered fixed differences appeared to show genes involved in serine-type endopeptidase or peptidase activity or inhibition, which in arthropods often play roles in immune system function, as well as digestion in Lepidoptera (Rodrigues Macedo *et al.* 2011). These genes may therefore be under selection if there are host plant differences between populations.

Scaffold	Gene Position	*H. erato* name	*Drosophila* name	Tblastx e-value	Function
0101	2,725,320 - 2,742,345	H*erato*0101.76	*Ir87a* (Not good match)	1.25329	Ligand-gated ion channel activity
0206	441,571 - 453,032	H*erato*0206.17	CG32700	1.24884e-05	No functional information
0206	435,428 - 448,903	H*erato*0206.18	No gene	0.000127646	-
0206	475,111 - 507,383	H*erato*0206.19	Repeat region	1.9445e-16	-
0606	1,173,097 - 1,182,493	H*erato*0606.39	CG32344	2.53055e-70	ATP binding; helicase activity; RNA binding
0801	2,772,552 - 2,808,142	H*erato*0801.48	*GstZ2* (CG9363)	1.40547e-102	Glutathione transferase activity
0901	7,843,452 - 7,853,199	H*erato*0901.259	CG17739	3.65004e-20	Serine-type endopeptidase inhibitor activity
1301	8,787,129 - 8,798,566	H*erato*1301.373	*Spn42Dd*	1.68981e-23	Serine-type endopeptidase inhibitor activity
1301	8,932,194 - 8,932,934	H*erato*1301.374	*hu li tai shao* (CG43443)	1.45421	actin binding
1601	196,230 - 297,213	H*erato*1601.12	*Vsx2* 2CG33980	7.0646e-50	sequence-specific DNA binding; transcription factor; homeobox
1601	303,064 - 311,105	H*erato*1601.14	No gene	0.225895	-
1910	2,108,239 - 2,121,003	H*erato*1910.130	*Cappuccino* (CG3399)	4.63036e-45	microtubule binding
1910	2,209,269 - 2,210,188	H*erato*1910.131	*Gpa2* (CG17878)	0.28073	G-protein coupled receptor binding
2101	11,895,505 - 11,968,709	H*erato*2101.397	CG4928	1.34386e-154	No functional information

Table 2.7 - Location and functional information (from *D. melanogaster*) of genes at or near clusters of fixed differences found between *H. demeter bouqueti* and *H. demeter beebei*.

2.4 DISCUSSION

The parallel hybrid zone across the Guianas is the largest phenotypic transition, in terms of species number, in the whole of the *Heliconius* radiation. Seven different species of *Heliconius* all exhibit the same colour pattern transition, with red/orange hindwing rays in the East but not in the West. This provides the perfect opportunity for testing the repeatability of evolution. The genetic control of these hindwing rays has already been mapped in two species found in this hybrid zone; *H. melpomene* and *H. erato* (Baxter *et al.* 2008b), to a single homologous locus. In contrast, very little genetic work, other than phylogenetic analysis, has otherwise been done for four of these seven species; *H. burneyi*, *H. xanthocles*, *H. demeter* and *H. aoede.* QTL mapping studies were beyond the scope of this study, given the difficulty in rearing these species, so I took a population genomics approach to find regions of divergence between individuals from either side of this hybrid zone, in the species *H. aoede* and *H. demeter*. Unfortunately samples of *H. burneyi* and *H. xanthocles* from this hybrid zone were not available. Regions of divergence found in this natural experiment should include the regions involved in colour pattern control. Interestingly, in both these species, the regions of divergence did include this same homologous locus *BD*, consistent with the hypothesis, that this region has repeatedly, across species, evolved a role in colour pattern control. However, perhaps surprisingly in both species these regions were not the regions showing the greatest divergence. With the greatest cluster of fixed differences found around an ommochrome pathway gene in *H. aoede*, and in *H. demeter* around a gene with unknown function (gene: H*erato*0206.17).

2.4.1 **Patterns of divergence**

Regions of divergence found across the Guianese hybrid zone should in theory include regions involved in colour pattern control, as this has been found to be true in other population genomic studies of *Heliconius* (Baxter *et al.* 2010; Counterman *et al.* 2010;

Nadeau *et al.* 2012, 2013, 2014; Supple *et al.* 2013). However, these previous studies have either used a biased, targeted approach, looking across complete tile paths or scaffolds of colour pattern loci and a few unlinked loci in *H. melpomene* or *H. erato*, or they have used a whole genome approach using RAD data and large sample sizes. In contrast, for *H. demeter* and *H. aoede* there are no species specific tile paths across known colour pattern loci, while only limited samples are also available. I did however have whole genome sequence libraries for these samples. I therefore used a test dataset of *H. melpomene* already sequenced from a hybrid zone in Peru to test this *de novo* genome approach. Across this Peruvian *H. melpomene* hybrid zone, both major colour pattern loci are known, while a reference genome for *H. melpomene* is also available. This means that I could test the success of this approach, where a *de novo* genome built from short reads with varying assembly parameters was used as the reference. To see if regions of divergence identified are the same as those found when using an actual reference quality genome.

Analyses using a *de novo* assembly as a reference returned approximately half the number of fixed differences to the analysis using the published reference quality genome. This is most likely due to incompleteness of reference and poorer mapping of reads. However, the signal was still very strong, with more than 80% of fixed differences typically found in the two colour pattern regions. In addition, approximately 7-8% of the remaining fixed differences found in another peak also unveiled by the *H. melpomene* reference genome analysis. This final peak is not known to control colour pattern but was not unexpected, as this locus has been found in GWAS studies between the same colour pattern races to be divergent (Nadeau *et al.* 2014). I also found these results to be robust to changes in the *de novo* genome assembly parameters, with all assemblies giving somewhat different, but comparable results (Table 2.1). This meant that if the optimal parameters for assembly did differ, for the *H. aoede* and *H. demeter* datasets relative to the *H. melpomene* test dataset, it is lilkely that the same main peaks of divergence should be found.

In comparison to the test dataset, where a very large proportion of fixed differences were found in just a few loci, the results from the analysis of the Guianese hybrid zone consistently showed a larger number of smaller clusters, with fixed differences spread between these. This is likely due to two factors. The first of these is that while in Peru the colour pattern differences are large, a postman pattern versus a dennis-rayed pattern, in the Guianas the phenotypic transition is more subtle, with only one element, the hindwing rays varying between the two colour pattern races. It is likely this leads to weaker selection across the Guianese hybrid zone. This stronger selection in Peru helps maintain a relatively narrow hybrid zone (Rosser *et al.* 2014). In contrast the hybrid zone across the Guianas is relatively wide and more variable. The sampling across the Guianas was therefore across a much greater geographic distance relative to the distance in Peru. It is likely this increases the noise to signal ratio due to the reduced homogenising effects of gene flow, with many of the smaller peaks found in the Guianas, perhaps due to drift rather than selection at colour pattern loci, as was generally shown though expanding sample sizes.

To determine the probability of finding a cluster of fixed differences of a given size in the genome, and therefore to assess the significance of clusters, I used a permutation method to simulate the distribution of fixed differences, this is similar to the bootstrapping methods used by others (Nadeau *et al.* 2012; Andrew & Rieseberg 2013). This method gave a crude size cut off below which a cluster was most likely random and non-significant. This random fixed difference cluster process essentially equates to drift. However, drift does not act upon each nucleotide independently to all others, but rather across sliding regions under linkage. Therefore these cut-offs, though useful, were likely conservative, and so in order to reduce noise I looked to expand the sample size across which these smaller clusters of differences were fixed, using short range PCR. This proved highly successful with all but one clusters of fixed differences quickly dropping out with just a few extra samples, suggesting that these clusters are due to allele frequency

differences, likely due to drift and therefore are unlikely to be involved in colour pattern control. In contrast, all five of the fixed differences from the second largest cluster in *H. aoede* overall (with 17 fixed differences), which were sequenced over larger sample sizes, were found to remain fixed, suggesting that this region is likely under selection across the hybrid zone.

2.4.2 *Cardinal*; ancient gene, novel function?

Ommochrome pigments have a conserved function across insect taxa, working as screening pigments that assist in the photoregeneration of rhodopsin, helping to tune the eyes of each species to the natural light conditions that they encounter (Stavenga 2002). These pigments are controlled by an array of conserved genes, which have generally been identified because of their associated *Drosophila* eye mutants (Haffter *et al.* 1996). These various eye mutant genes can be split into three main groups, Granule genes, Pigment synthesis genes and ABC transporter genes (Shoup 1966; Haffter *et al.* 1996; Reed & Nagy 2005). Each of these performs a different function in the eye of the fly, with ABC transporter genes first transferring the pigment pre-cursors across the cell membrane, where pigment synthesis genes that code for different enzymes produce the pigments. These are then moved to the pigment granules, whose biogenesis are controlled by an array of granule genes. Together these genes produce and control the pigmentation in the eyes of *Drosophila*.

The orange and red pigments that pattern the wings of *Heliconius* butterflies are also ommochrome pigments, respectively called Xanthommatin and Dihydro-xanthommatin, while the yellow precursor to these is the pigment 3-Hydroxy-kynurenine (Gilbert 2002; Reed & Nagy 2005; Reed *et al.* 2008). In *Heliconius* and other butterflies many of the genes first identified in the eyes of *Drosophila* have now been found to be expressed during wing development (Reed & Nagy 2005; Reed *et al.* 2008; Ferguson *et al.* 2011; Hines *et al.* 2012). However, QTL mapping studies in *Heliconius,* looking at the segregation of different

colour pattern elements, have not found these genes to be linked to changes in these phenotypes (Joron *et al.* 2006a). For the hindwing rays of *H. melpomene* and *H. erato*, the gene *optix* (discussed in next section) has instead been found to be associated with these changes (Baxter *et al.* 2008b; Wallbank *et al.* 2016), and its expression correlated during development with red pigmentation too (Reed *et al.* 2011; Martin *et al.* 2014b). While not an ommochrome pathway gene, in its role as a transcription factor *optix* is thought to control a barrage of downstream genes including pigment enzymes like *cinnabar* and *ebony* (Martin *et al.* 2014b; Merrill *et al.* 2015).

In the results presented here, it is striking that over twenty percent of overall fixed differences and the greatest concentration of fixed differences across the *H. aoede* genome, are in and around a gene that is in *Drosophila* associated with its own eye mutant. This ommochrome pathway gene is *cardinal*. Mutations at this gene have been found to block pigmentation of the secondary pigment cells in eyes, while causing excessive pigmentation of primary pigment cells (Stark *et al.* 1981; Tearle 1991). In addition, the temperature sensitive period of mutant *cardinal* alleles coincide with the onset of eye pigmentation (Tearle 1991). Overall this gene makes an intriguing candidate for a gene involved in the pigmentation and patterning of wings, given its conserved function in ommochrome pathways.

The argument for this possible role is further supported by recent work on a c*ardinal* mutant in the silkmoth *Bomyx mori*. This mutant has white eggs and pink-eyes, and lacks red pigmentation on the epidermis of final Instar larvae (Osanai-Futahashi *et al.* 2016). In concert, at a cellular level 3-hydroxykynurenine accumulates relative to the wildtype, suggesting a fault in the conversion of this yellow pigment to the orange and red Xanthommatin and Dihydro-xanthommatin (Osanai-Futahashi *et al.* 2016). Given this conserved role of *cardinal* in Xanthommatin and Dihydro-xanthommatin biosynthesis, it seems plausible that these two highly divergent alleles found on either side of the

Guianese hybrid zone, and which contain *cardinal*, may play an important role in the loss of the rays and Dihydro-xanthommatin pigmentation on the hindwings of *H. aoede centurius*.

This would be the first example in *Heliconius* where changes in red patterning have been found not to be controlled solely by regulatory changes in o*ptix* expression alone. While *optix* expression patterns correlated with red patterning have been found across *Heliconius* taxa (Martin e*t al.* 2014b), this only actually implies that o*ptix* plays an important conserved role across taxa, but does not necessarily mean that the loss of a certain element can only be achieved through a loss in o*ptix* expression. Actual genetic studies looking at the genetic basis of these convergent phenotypes have in contrast to expression studies, tended to have a much narrower taxanomic focus, with QTL mapping and population genomic approaches only really applied to *H. erato*, *H. melpomene* and some of their close relatives (in the Silvaniforms, and *H. cydno* and *H. himera*).

In light of the results presented here, it seems plausible that regulatory changes in developmentally downstream genes like c*ardinal* are also able to achieve a similar phenotypic result to changes in *optix* expression. Around c*ardinal* the majority of fixed differences were found within introns rather than within the exons. This is consistent with a model in which regulatory changes at c*ardinal*, rather than protein coding sequence changes, are leading to loss of red pigmentation in one part of the wing, while being maintained in other parts. In *Heliconius* where evolution has been assumed to be convergent based on this limited taxonomic sampling (Baxter e*t al.* 2008b; Reed e*t al.* 2011), this is perhaps an important example of how evolution can be more flexible, and can arrive at similar phenotypes through multiple evolutionary solutions, that involve changes to genes in the same developmental pathways.

2.4.3 **Repeated evolution at the *rays* locus**

Although the divergence found around *cardinal* in *H. aoede* is striking, in both *H. demeter* and *H. aoede* fixed differences were also found near the gene *optix* known to control the rays phenotypes in *H. melpoemene* and *H. erato* as well as other species of *Heliconius* (Reed *et al.* 2011; Martin *et al.* 2014b). Furthermore, the fixed differences found in *H. aoede* and in *H. demeter* were not just near *optix*, but in fact close to the *cis*-regulatory module thought to actually control the rays phenotypes in *H. erato* (Van Belleghem *et al.* 2016) and *H. melpomene* (Wallbank *et al.* 2016). The evolution of *cis*-regulatory modules like this have often been found to be the main driving force behind much rapid morphological evolution (Wittkopp & Kalay 2012) as through the evolution of novel enhancers, genes and developmental pathways can either be co-opted, or assembled *de novo* into new pathways, for novel functions, while the function of these genes can be conserved across other developmental networks (Monteiro & Podlaha 2009).

It therefore seems likely that o*ptix* expression plays a role in the patterning of rays in both *H. aoede* and *H. demeter*. However, given that both c*ardinal* and *optix* are plausible candidates for the control of rays in *H. aoede*, the results from this analysis are hard to interpret. If both of these loci are indeed involved in colour pattern controls then hypotheses are possible; i) that only one of these loci controls the loss of rays, while the other may be a modifier for some other colour patterning phenotype, or ii) that both loci work epistatically to control the loss of rays across the Guianas. Thus second hypothesis could work with either both able to switch on or off the rays, or one working as a modifier of the main switch locus.

2.4.4 **Conservation across subspecies**

An analysis including an outgroup sample from Peru was also carried out for both *H. demeter* and *H. aoede*, with fixed differences looked for between two groups; a group

composed of Surinamese non-rayed samples, and a second group composed of rayed samples from and French Guiana and Peru. This analysis was designed to test if fixed differences found between Surinamese and French Guiana, were fixed across allopatric populations that shared the rayed colour pattern phenotype. This was found to be true for the *H. aoede* analysis with both the *cardinal* and *BD* region showing a number of fixed differences with the Peruvian outgroup sample included. This suggests that the alleles at both loci, found in Peru and French Guiana, are more similar to each other than they are to that of the Surinamese allele, and supports the hypothesis that both are involved in colour pattern control, with some of these fixed SNPs possibly functional. In contrast, the majority of fixed differences in *H. demeter* dropped out, with only one fixed difference at the *BD* locus. This suggests that the genetic control of the rays is not conserved across rayed subspecies and that the fixed differences found are not functional.

2.4.5 **Serine proteases**

Despite many of these smaller peaks dropping out with increased sample sizes, all clusters of fixed differences were still blasted to *Heliconius* reference genomes and the *Drosophila melanogaster* genome in order to identify possible functions for genes showing elevated levels of divergence. Serine protease and serine protease homolog genes do form a large family in insects with ~100 known in the plant hopper, *Nilaparvata lugens* (Bao *et al.* 2014), and ~200 known from *D. melanogaster* (Ross *et al.* 2003), while serpin genes (Serine Protease Inhibitors) form a somewhat smaller family with ~30 genes known (Reichhart *et al.* 2011). These serine-type endopeptidase or serine peptidase activity genes are known to dominate the larval gut environment and have been found to contribute to about 95 % of the total digestive activity in Lepidoptera (Rodrigues Macedo *et al.* 2011), while also playing an important role in insect immunity (Zou *et al.* 2006). Given these numbers of serine protease pathway associated genes present in insect genomes, it was striking that out of the 12,699 predicted genes in *H. melanogaster* (Dasmahapatra *et al.* 2012), in both

demeter and *H. aoede* genes involved serine peptidase activity were found to show elevated divergence. This elevated divergence in a gene involved in serine peptidase activity can also be seen across a *H. erato* hybrid zone (Nadeau *et al.* 2014). It seems possible that this could be caused by selection due to host plant differences between populations of either sides of the hybrid zone.

In *H. aoede,* the serine peptidase activity genes found were actually around the location of the second largest cluster of fixed differences in *H. aoede* overall, with 17 fixed differences on scaffold HE671576. These differences also remained fixed over larger sample sizes, and alternatively could play a role in colour patterning. In *D. melanogaster* the genes in this region; *easter* (CG4920) and *Spatzle-Processing Enzyme* (CG16705) both have described roles in immunity, but in addition play a role in ventral-dorsal patterning in the egg and embryo, with the *easter* protease processing the *pro-Spatzle* protein to generate the Toll ligand during development (Jang *et al.* 2006). Mutations disrupting this process in *Drosophila* are known to cause changes to embryonic cuticle patterns (Jin & Anderson 1990). In butterflies many of the genes known to play an important role in colour patterning, have other deeply conserved and important roles in development as homeobox genes. These play numerous important roles including anterior/posterior axis specification (*hedgehog*), or Proximal/distal pattern formation (*distal-less*) (Brakefield 1998; Brunetti *et al.* 2001; Taylor 2002). It is possible therefore that these fixed differences found around *easter* (CG4920) and *Spatzle-Processing Enzyme* (CG16705) may play some role in colour pattern formation in *H. aoede*.

2.4.6 **Conclusion**

Across the rayed, non-rayed Guianese hybrid zone in *H. demeter* and *H. aoede*. Islands of divergence were found in homologous sequence in both species close to the region identified as containing the *rays* module in *H. melpomene*. However, more surprisingly, I found other much larger islands of divergence unique to each species at other loci. In *H.*

demeter the function of the gene around this region is unknown, while it is possible that this elevated divergence might be caused by an inversion as is the case in this genomic region in *H. erato*. However, in *H. aoede* the region of greatest divergence was an ommochrome signalling pathway gene, *cardinal*; an excellent candidate for a gene involved in colour patterning. This suggests that the Guianese colour pattern shift in at least *H. aoede* may not solely be determined by regulatory changes in *optix* expression, as was previously thought to be the case across *Heliconius*, given the striking genetic convergence between *H. melpomene* and *H. erato*.

3. The genetics of diversity: the dennis-rayed mimicry ring of *H. melpomene*.

3.1 INTRODUCTION

Biological diversity exists at many scales, from diversity at higher taxa, species diversity, and intraspecific diversity, and in terms of both the phenotype and the genotype. However, in order to understand the origins of diversity at the species level and above, one must also understand the interactions of diversity at the intraspecific level and the processes that drive and determine this diversity. The mimicry rings of *Heliconius* butterflies provide a perfect system for exploring intraspecific diversity (Mallet & Joron 1999). These butterflies possess bright aposematic colour patterns, and form Müllerian mimicry rings in which different species share colour patterns, and thus the costs of predation as well as the benefits of protection that their shared colour patterns confer (Merrill *et al.* 2015). Paradoxically, as well as striking convergence between species, *Heliconius* also show great diversity within species (Joron & Mallet 1998). Two of these species, *H. melpomene* and *H. erato* that diverged ~10-12 million years ago (Kozak *et al.* 2015), are found across much of the neotropics, with approximately 40 different subspecies, each with a different colour pattern (Hines *et al.* 2011). These can be split into two main mimicry rings on either sides of the Andes, with postman butterflies in central America and the western coastal side, and the dennis-rayed butterflies of lowland Amazonia on the eastern side (Hoyal Cuthill & Charleston 2012).

Understanding the genetics of this colour pattern diversity started with the work of Turner and Crane (1962) and Sheppard (1985), conducting crossing experiments between divergent forms in order to understand the segregation of these traits and their genetic control. This established that much of the diversity in red-orange and yellow elements was determined by a small number of major effect loci, that act as colour pattern switches. More recent work has now mapped these loci to regions of the genome, and has revealed that these major effect loci are in fact homologous between the two species (Baxter *et al.* 2008b).

Probably the most well understood of these major effect loci is the *BD* locus, that controls the main red-orange elements, like dennis patches, red forewing bands, and hindwing rays (Figure 3.1a) in *H. melpomene* and *H. erato* (Sheppard *et al.* 1985). This locus has been mapped to chromosome 18 in both species (Baxter *et al.* 2008b; Papa *et al.* 2013), as well as in *H. hecale* (Huber *et al.* 2015) The gene at the heart of these major effects is the transcription factor *optix*, which has been found to be expressed during development in red regions of the wing just prior to ommochrome pigmentation(Reed *et al.* 2011). This expression pattern been confirmed in both *H. erato* and *H. melpomene* as well as a range of *Heliconius* species quite closely related to the latter (Reed *et al.* 2011; Martin *et al.* 2014b). Population genomics studies also support a role for it in other species quite closely related to *H. melpomene*, in some cases implicating a role for introgression in spreading the effects of this locus throughout the genus (Chamberlain *et al.* 2011; Dasmahapatra *et al.* 2012; Pardo-Diaz *et al.* 2012). On the basis of this assumption, more recent work using the diversity of recombinant wing pattern forms across species, appears to have identified some of the *cis*-regulatory modules controlling this *optix* expression (Wallbank *et al.* 2016).

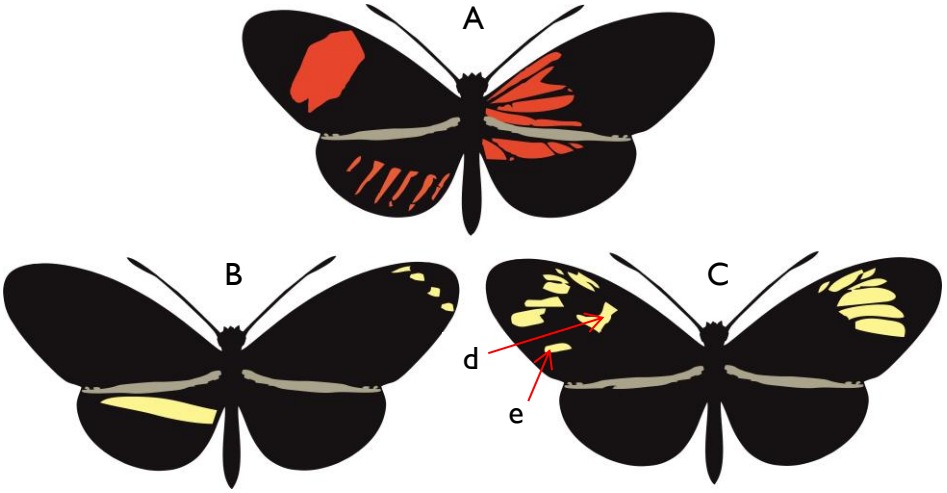


Figure 3.1 – Colour pattern elements controlled by the three major wing pattern loci in *H. melpomene* and *H. erato*. A) Shows elements controlled by *BD* locus. On the left, red postman forewing band, and hindwing rays; on the right, hindwing and forewing dennis patches. B) Shows elements controlled by the *Yb/N* locus. Left shows effect as *Yb*, controlling hindwing yellow bar found in some postman forms; right, shows effect as *N* controlling apical band in *H. hecale*. C) Forewing bands variants in the dennis-rayed mimicry ring, left shows broken band, right shows medial band only. *Ac* locus has been shown to control presence and absence of (d) dumbbell/cell spot and (e) belem spot, in *H. erato*.

While *BD* controls the diversity of red-orange colour pattern elements, two other major effect loci, *Yb*/*N* and *Ac*, control many of the possible yellow colour pattern elements (Figure 3.1b). The first of these, *Yb*/*N,* controls both the hindwing yellow bar, and together with the *BD* locus controls the switch from a red to yellow forewing band (Sheppard *et al.* 1985). Further, in *H. hecale* this locus has been shown to control the apical forewing band (Huber *et al.* 2015), while in *H. melpomene*/*H. cydno* under the name of *Sb* (Linares 1996; Jiggins *et al.* 2005; Ferguson *et al.* 2010) and in *H. erato* under the name of *Cr* (Jiggins & McMillan 1997; Kronforst *et al.* 2006a) it has been found to control the white hindwing margin. This locus has been found to be homologous across not just these species, but is also with the supergene *P* that controls all of patterning in *H. numata* (Joron *et al.* 2006b; Jones *et al.* 2012). One of the genes found at this locus to be involved in determining aspects of this colour pattern variation is the gene *cortex,* this has been found to be divergent between races of *Heliconius* with different *Yb*/*N* phenotypes, as well

as showing differential gene expression between black and yellow wing regions during development (Nadeau *et al.* 2016). Furthermore this gene is not just implicated in colour patterning in *Heliconius* species, but also in the silk moth *Bombyx mori* and the peppered moth *Biston betularia* and *Bicyclus anyanna* (Nadeau *et al.* 2016; Hof *et al.* 2016)

The other locus involved in patterning the forewing band elements, *Ac*, is proposed to do so by controlling the distribution of melanised scales across the butterflies wings. This was first mapped to chromosome 10 in crosses between the two species *H. cydno* and *H. pachinus* and was found to control the presence or absence of melanic scales on the proximal regions of both the fore and hind wings (Kronforst *et al.* 2006a). Crosses have also shown that this same locus affects forewing band shape variation in *H. erato* and controls the presence and absence of the broken band in *H. erato* (locus called *Sd* in *H. erato*) (Martin *et al.* 2012; Papa *et al.* 2013). In *H. melpomene*, QTL mapping has shown that this locus *Ac* is also linked to phenotypic variation in forewing band shape (Martin *et al.* 2012). In addition, insitu hybridisation studies and work with Heparin injections, an analog of an extracellular matrix compound that expands the gradients of morphogens, supports the theory that the gene *WntA,* a morphogen*,* is the most likely gene controlling melanisation in both of these species (Martin *et al.* 2012; Gallant *et al.* 2014b; Kronforst & Papa 2015).

Although these major effect loci have been found to control much of the diversity of *Heliconius* wing colour patterning, evidence for a large number of other smaller effect loci has also been found (Baxter *et al.* 2008a; Papa *et al.* 2013; Nadeau *et al.* 2014; Huber *et al.* 2015). Two of these, on chromosomes two and seven are known only to effect the red forewing band size and shape in *H. melpomene* (Baxter *et al.* 2008a). Another on chromosome seventeen has been found to be associated hindwing yellow bar in *H. erato* (Nadeau *et al.* 2014), while a locus on chromosome thirteen has been implicated in the rounding of the yellow forewing band in *H .erato* (Nadeau *et al.* 2014) and in red forewing

band shape in *H. melpomene* (Baxter *et al.* 2008a). In addition, a number of other minor effect loci, have been identified across other linkage groups, explaining variation in the number of red and white scales, and shape of the forewing bands of *H. erato notabilis* (Papa *et al.* 2013). Furthermore, in chapter 2 I identify two different loci, one in each of *H. aoede* and *H. demeter* that are divergent between colour pattern races from a hybrid zone across which the hindwing rays phenotype varies, of which at least one seems likely to be involved in colour pattern control.

Despite the number of mapping crosses in *H. melpomene*, so far all featuring an Amazonian dennis-rayed individual have seen this crossed with an individual from a coastal postman population. In order to better understand the effects of these minor effect loci in the Amazonian dennis-rayed mimicry ring, I took advantage of the availability of divergent stocks of *H. melpomene aglaope* from Amazonian Peru and *H. melpomene meriana* from Suriname (Figure 3.2) to identify minor effect loci involved in the control of forewing band shape variation in the dennis-rayed mimicry ring. The *Ac* locus, known to control variation in both red and yellow forewing band shapes, has been mapped in *H. melpomene* (Martin *et al.* 2012), and has been shown to control the full broken band phenotype in *H. erato* (Papa *et al.* 2013) and the cell spot found in *H. cydno* (Kronforst *et al.* 2006a). However, in *H. melpomene* mapping crosses have not been conducted with this broken band phenotype. While *Ac* seems a likely candidate, the confirmation of this is of value, and additional modifier loci may well be involved. The crosses described in this chapter provide an opportunity to do this, as well as to refine the *Ac* locus to a smaller region in *H. melpomene*, and help confirm the results from Heparin injections that suggest *WntA*'s involvement in melanin patterning.

While the switch between white and yellow pigmentation has been shown to be controlled by a locus called *K* on chromosome 1 (Kapan 1998; Kronforst *et al.* 2006b; Huber *et al.* 2015), the switch between red and orange pigmentation of colour pattern

elements controlled by *BD* has so far not been mapped. This switch happens only in *H. erato* and *H. melpomene*. In general, orange pigmentation in the form of the ommochrome xanthommatin (Joron *et al.* 2006a) is found in the dennis rayed subspecies. This pigment is also found in their comimics and in the silvaniform species (Brown 1976), while red pigmentation in the form of dihydroxanthommatin (Reed *et al.* 2008) is found in the forewing bands of most postman races in the Guianas and Central America (Sheppard *et al.* 1985). However in the Guianas a population with red pigmented dennis-rayed elements is found. The existence of this population; combined with crosses, has shown that the loci controlling red-orange pigment and red-orange element patterning are unlinked. This locus controlling this pigment change has been previously termed *Or* (Sheppard *et al.* 1985) and can be mapped in the experimental design used in this chapter.

Figure 3.2 – Example of variation segregating in the F2 mapping family B10. Top row, shows ventral wing surfaces from pure subspecies grandparents; middle row, shows dorsal and ventral surfaces of F1 parents; and bottom row shows segregating phenotypes in the F2 progeny on ventral surfaces.

3.2 METHODOLOGY

3.2.1 **Crossing experiments**

Stocks of *H. melpomene meriana* (from western Suriname) and *H. melpomene aglaope (*from Amazonian Peru) were started from wild caught individuals. Both of these colour pattern races are from the dennis-rayed mimicry ring found throughout the Amazon. *H. melpomene meriana* were collected from Victoria, Suriname (5.113892 N -54.990106 W), *H. melpomene aglaope* were collected from Shucushyacu, Peru (-6.007558 S -75.884416 W).

F2 and backcross mapping families were generated from these stocks in the insectaries at York University. Butterflies were kept in cages measuring 1.2m (Length) x 1.5m (Width) x 2.5m (Height), and fed on a mixture of honey, pollen and water. Larvae were fed on a variety of *Passiflora* species, with *P. caerulea* the main feed plant for *H. melpomene meriana* and *P. stipulata* the main food plant for *H. melpomene aglaope.* Eggs were laid on shoots kept fresh in water, larvae were kept on these while young and then moved to be reared in individual plastic pots from second instar to emergence.

3.2.2 **Sample preservation and sequencing**

Upon emergence wings from F2 and back cross progeny were removed and phenotypes, whole bodies were preserved in dimethyl sulfoxide (DMSO) salt solution (20% DMSO, 0.25 M EDTA, saturated with NaCl) at -20°C. Mapping family fathers were preserved directly after mating while and family mothers were taken once they had died naturally. These family parents were again stored in DMSO salt solution at -20 °C. RNA-free genomic DNA was extracted to a concentration of approximately 15ng/µl from thoracic tissue using a Qiagen DNeasy Blood and Tissue Kit following the standard protocol provided by the manufacturer. Restriction site Associated DNA (RAD) libraries were prepared (by K. Dasmahaptra) using a modified protocol from Etter *et al* (2011), using a

PstI restriction enzyme, sixteen 6bp P1 barcodes and eight indexes. DNA was covaris sheared to 300-700bp and gel size selected. 128 individuals were sequenced per lane, with 125bp paired end reads, on an Illumina HiSeq 2500.

3.2.3 Segregation of phenotypic variation

Both the ventral and dorsal sides of butterfly wings were scanned 1-7 days after emergence using a Canon LiDE 700F scanner (with the MP Navigator EX 2.1 driver, under the colour document setting and with 300dpi resolution) in order to be phenotyped. Hindwing and forewing dennis elements were present in both wildtype parental phenotypes and were not of interest in this study. Therefore of the red/orange elements, only the presence or absence of rays was necessary to record. Two aspects of the forewing band were scored; these were the presence and absence of the Dumbbell and Belem spots. Three scores were possible; 0 for complete absence (as is found in *H. melpomene aglaope*), 1 for presence (as is found in *H. melpomene meriana*) and 0.5 when these phenotypic characteristics were partially present. Chi squared tests were implemented in R v3.3.1 to test for deviations from the expected ratios for a recessive phenotype (broken) controlled by a single Mendelian locus. Two thresholds were used with bands scored using a relaxed threshold of >=2.5 (collated across both the dumbbell and belem spots) for the presence of the broken band, and a more stringent threshold for bands to be scored as broken in which the dumbbell spot was scored as completely present (2) and the belem spot was scored as >= 1.5. This enabled us to gauge variation in this trait between mapping families to help select families for use in QTL mapping analysis.

Red and Orange colouration was recorded using Corel Photopaint X6 by recording the mean RGB values in a 5x5 pixel point centrally located in the dennis patch of the forewing. Measurements were taken for both dorsal and ventral sides. Principal component analysis was then carried out between parental strains. This was implemented

by first adding one (to account for 0s) and then log_{10} transforming RGB values from both the dorsal and ventral surfaces, to account for a moderate positive skew. Data was then centred by subtracting column means, before singular value decomposition was carried out with svd() on the covariance matrix. In R v3.3.1 eigenvectors from this principal component analysis were used to transform additional F1, F2 and backcross progeny that were then added to plots. This enabled the calculation and visualisation of F1, F2 and backcross progeny on principal component axis describing variation between the two species, in order to explore the segregation of parental pigmentation in progeny.

3.2.4 **Forewing band shape**

Scanned images of butterflies were first brightened in Adobe Lightroom 5 for easier visualisation. 'Curves' around each part of the forewing band were then traced in tpsDig2 (Rohlf 2013a) on the ventral surface. This band was generally composed of seven distinct elements of varying size. The number of points for each curve was as follows: Curve one, 11; curve two, 20; curve three, 10; curve four, 20; curve five, 25; curve six, 35; curve seven, 25. The option '*resample by length'* was then used in order to equally space points around each curve. In order to convert curve points to semi-landmarks, the 'append tps curves to landmarks' function was first used in tpsUtil (Rohlf 2013b). This landmarks file was then opened in the 'make sliders file'. This function allows the conversion of landmarks to semilandmarks, with the central landmark of each triplet, enabled to slide parallel to the difference between the two landmarks directly either side of it. This removes tangential variation so that points along the outline curve match as well as possible to the positions of the points on the reference configuration (Perez *et al.* 2006) by minimising bending energy during Generalised Procrustes Superimposition (Rohlf 2013b). If any of the seven elements of the main forewing band were missing this was accounted for accounted for by drawing a curve with the full number of points

maintained. These were then modified after conversion to landmarks in tpsDig2, by compiling the landmarks onto one single point.

Once curves had been converted to landmarks, Generalised Procrustes Superimposition was carried out using the gpagen command from the R package geomorph. Following this, principal component analysis was again carried out between 10 samples from each parental strain. Data was centred by subtracting X and Y means for each landmark from the data, before singular value decomposition was carried out with svd() on the covariance matrix. This was done in R v3.3.1. Eigenvectors from this PC analysis were used to transform additional F1, F2 and backcross progeny from three families, these were then added to plots. This enabled the calculation and visualisation of F1, F2 and backcross progeny on principal component axis describing variation between the two species, in order to explore the segregation of this multivariate trait.

3.2.5 **Linkage map construction**

Each RAD library of 15-16 individuals was first processed using the process radtags from Stacks (Emerson *et al.* 2010) in order to split each individual by barcode sequence into separate forward and reverse fastq files. Following this, read group information, machine number and read pair, was added back to the newly processed fastq files. BWA mem (Li & Durbin 2009) and SAMtools view for BAM conversion (Li *et al.* 2009) were then used to map the reads of each individual against the reference *H. melpomene* genome v2 (Davey *et al.* 2016). BAM files were subsequently sorted with SAMtools, and PCR duplicates marked with Picard-tools v1.1 MarkDuplicates (broadinstitute.github.io/picard/). At this point a custom python script from John Davey sex_by_coverage.py was used to check BAM files for Z vs autosome coverage. This was done in order, to check that the pattern of males and females unique to each library according to the sequence data corresponded to that expected. In this same way any possible errors during library prep could be detected. No such errors were found.

HaplotypeCaller from the GATK v3.4-46 (McKenna *et al.* 2010) was then used to for variant calling, with heterozygosity set to 0.001 and minimum pruning set to 2. This VCF file was then converted to a variants table using VariantsToTable from the GATK v3.4-46, and filtered to a file with genotype calls using a Perl script (Appendix 20). Genotypes with > 150x coverage, < 5x coverage, genotype quality (GQ) less than 20, SNP quality less than 30 or mapping quality less than 20 were ignored. In this way low quality genotypes were set to missing, with a GQ of 20 equating to a genotype that is estimated to having a likelihood 100x more than that of the second most likely genotype. A Perl script was then used to filter out markers with more than 20% missing data, and to estimate the missing data for each sample (Appendix 21).

The genetic linkage map was built using a combination of modules from both LepMAP2 (Rastas *et al.* 2016) and LepMAP3 (https://sourceforge.net/projects/lep-map3/). The first step in map construction was to convert this filtered VCF file containing only these high quality markers, to a posteriors file, and then a linkage file, using scripts bundled with the LepMAP programs. To this a pedigree was added, and checked by calculating identity by descent (IBD) between samples using plink1.9 (https://www.cog-genomics.org/plink2; Purcell *et al.*, 2007; Chang *et al.*, 2015). Three progeny showing a lower than expected Pi score when compared to their parents were removed at this stage, as a low IBD score indicates they had been incorrectly assigned to this family, and their inclusion could interfere with linkage map construction and QTL mapping.

Parental genotypes were then called and corrected using the ParentCall module from Lep-MAP2, with non-informative markers set to be removed, and a Zlimit of 5. This ascribes markers as having Z inheritance if they meet the required log-odds difference. Markers were then filtered using the Filtering2 module from Lep-MAP3, with dataTolerance of 0.01. This sets the significance limit for segregation distortion. Mapping families were then split into separate files so that the module SeparateIdenticals could be

run. lodLimit options were set to 20 for maternal markers, log10 2^(n-(n/10)) for paternal markers and log10 3^(n-(n/10)) for intercross markers (n = number of individuals in the cross; calculation based on 2 possible genotypes for paternal markers, 3 for intercross markers, and allowing for 10% missing individuals). The additional options were also set; betweenSameType to1, lod3Mode to 2 and keepRate to 1. The Lep-MAP3 module OutputData was then used to adjust the genotype posteriors file, so that identical markers were set to have exactly matching segregation, again lod3Mode was set to 2 while sizeLimit was set to 3. These posterior files for each mapping family were then combined, so that the Lep-MAP3 module SeparateChromosomes2 could be run. LodLims between 5 and 15 were tested empirically, with a lodLimit of 10 and sizeLimit of 200 eventually chosen. This recovered 21 linkage groups, the known haploid number of chromosomes in *Heliconius melpomene*.

These linkage assigned markers were then ordered with the Lep-MAP2 module OrderMarkers, with initial recombination set to 0.05 for males and 0 for females to reflect achiasmatic recombination in Lepidoptera. Both male informative markers (heterozygote in the father) and dual informative markers were used (heterozygote in the father and mother) by setting informativeMask to 1 and 3. OrderMarkers, also estimates error scores for each marker, markers with error score > 0.1 were then removed from the linkage map, along with any markers producing large gaps, these were usually found at the ends of the linkage maps. Having removed these markers from each linkage group map, a Perl script was then used to remove all but the markers remaining in each linkage group from the map file (Appendix 22). This process results in a set of maps constructed from high quality markers, but with poor coverage across each chromosome. In order to extend coverage across each linkage group the Lep-MAP3 module JoinSingles2 was used, with lodLimit set to 40 and lodDifference set to 10. Again, informativeMask was set to 1 and 3, while lod3Mode was set to 3.

Following JoinSingles2, OrderMarkers was run again with the same initial recombination parameters, but with informativeMask set to 1 for most linkage groups, so that only paternally informative markers were included, and with minimum error set to 0.01. However, informativeMask was set to 1 and 3, for linkage groups 10 17 and 19 (Hmel2 chromosomes 15, 14, and 3 respectively) due to a paucity of markers. After this, linkage groups were refined. Markers were removed using Perl scripts (Appendix 23) for three possible reasons; 1) if their error rate was now greater than 0.02, 2) if they were found to have long gaps to the nearest markers, or 3) if they belonged to a Hmel2 chromosome different to that of the majority of markers on that linkage group. With these markers removed the marker order was re-evaluated with OrderMarkers with improve order set to 1.

All markers informative in both mapping families were then used as a basis for the final map. For some linkage groups this alone was sufficient. However, if these markers did not cover parts of a linkage group, markers paternally informative in one mapping family but not the other were also included in the maps. Again, markers were discarded if there placement did not make sense given the placement of other markers and the expected Hmel2 genome order. The AchiasmaticMeiosis module from Lep-Map2 was then used to convert all markers into paternally informative markers. Again, marker order was re-evaluated with OrderMarkers with improve order set to 1, and InformativeMask set to 1. In most cases, the markers that had been paternally informative only in one mapping family were now seen to be paternally informative in both, and ordered accordingly, to give the final 21 linkage groups for QTL analysis. Marker names give *H. melpomene* genome v2 scaffold and position.

3.2.6 **QTL analysis**

QTL analysis was carried out using R/qtl for univariate traits and with a combination of R/qtl and R/shapeQTL for multivariate traits. Principal component analyses for

multivariate traits were implemented with the prcomp() function in R. For all traits jittermap from R/qtl was first used to adjust the positions of markers in the linkage map that had been assigned to the same centiMorgan positions. Markers showing segregation distortion were then removed from all mapping families if they had a $-\log_{10}p > 15$ from chi-squared tests of Mendelian segregation in any family. Genotype probabilities were calculated separately for each family using a step size of 1cM and the Haldane mapping function, these families could then be combined using the c.cross() function, with family included as a covariate during QTL scans. This controls for each family having a different mean value. Genome wide scans with scanone() from R/qtl were then carried out in univariate analyses, while scanoneShape() from R/shapeQTL with a 'Pillai' test was used in multivariate analyses (with results used from the additive model). Significance for each analysis was then estimated using 1000 permutations (unless specifically detailed otherwise), also implemented using scanone() and scanoneShape(). scanoneShape() could not incorporate the Z chromosome, and so it should be noted that this has been dropped from multivariate analyses. In these the trait together with any covariate (family) is reordered across individuals, while original genotype probabilities are kept constant (Churchill & Doerge 1994). The size of each QTL was estimated using Bayesian 95% confidence intervals through the bayesint() function from R/qtl, and where LOD was high enough with LOD 1.5 drop-off intervals using the lodint() function. For univariate traits QTLs were further refined using the refineqtl() function from R/qtl. Specific QTL models were also fit to the data to further test the statistical significance of QTLs and to estimate effect sizes. For univariate traits this was done with makeqtl() and fitqtl() from R/qtl, while it was done with stepwiseqtlShape() from R/shapeQTL for multivariate traits (using an additive model only).

Two aspects of the forewing band were scored; these were simply the presence and absence of the Dumbbell and Belem spots (Sheppard *et al.* 1985). The presence and absence of these were scored on both the dorsal and ventral sides of the wings. Three

scores were possible for either side; zero for complete absence (as found in *H. melpomene aglaope*), one for presence (as found in *H. melpomene meriana*) and 0.5 when these phenotypic characteristics were partially present. An overall score of four therefore indicated complete presence of both elements, across both the ventral and dorsal surface of the wings, while a score of zero indicated complete absence. For QTL mapping analysis the broken band was treated as a single binary trait, with the relaxed threshold for scoring the broken band used, whereby a broken band presence and absence is scored using a threshold of >= 2.5. QTL analysis was carried out separately for each mapping family, as well as for both combined with mapping family used as a covariate. All analyses were implemented using Haley-Knott regression and a binary trait model. In order to assess effect sizes of each QTL an additive model including family as a covariate was fit to the data.

In order to identify QTLs involved in the differences in red-orange wing pigmentation log10 transformed RGB values from both the dorsal and ventral forewing surfaces were first taken for all progeny where possible in mapping families. Principal component analysis was then carried out using the prcomp() function in R v3.3.1. This confirmed the major RGB variables contributing to variation in wing pigmentation. A genome wide scan was then carried out using Haley-Knott regression, with a normally distributed trait model in Rqtl, on the single main RGB variable in wing pigmentation. Permutations (3000x) were then used to assess the significance of identified peaks. In addition to QTL mapping using a normal model on the $\log_{10}$ transformed ventral green RGB scores, multivariate QTL analysis was also carried out on principal component analysis of all transformed RGB values, from both dorsal and ventral wing surfaces. This second method has the advantage of being able to incorporate more variation that might also be important in describing red-orange pigmentation variation. This was done on all phenotyped individuals from sequenced families, as well as on each sequenced family separately.

QTL analysis of medial band shape was carried out on principal component analyses of both the combined families together, and of each family separately, using the same phenotyping protocol as used in the analysis of phenotypic segregation. All Principal component axes found to describe > 1% of variation were used in Genome wide scans of LOD. Log transformed centroid sizes were included as a covariate in these genome scans, while family was also included in the combined family analysis. Effects on medial band shape of QTLs were calculated using fitqtl() using AA and AB genotypes, and plotted with plot.shapeEffect() from R/shapeQTL. This was done for each family separately, due to differences in mean shape between the mapping families. Plots were done for each QTL identified in that family, with QTL positions given as the marker identified as having the highest LOD score on that chromosome in that family.

3.3 RESULTS

3.3.1 **Segregation of phenotypic variation**

In total I obtained five F2 intercross families and two backcross (to *H. melpomene meriana*) families (Table 3.1), from F1s from three parental strain matings. Intercross families B10 and B13 were started from two pairs of F1 siblings all from the same F1 family. Intercross families B11, B12, were also all related, coming from two pairs of F1 siblings from a different F1 family, these intercross mapping families were also related to families B8 and B14, respectively made by crossing a female F1 from the same family, back to a stock male of *H. melpomene meriana*, and a male F1 back to a stock female of *H. melpomene meriana*. The intercross family B5 was unrelated to all other families.

F1s were all found to have hindwing rays (Figure 3.2). This phenotype did not segregate significantly differently from that expected from Mendelian ratios for F2 intercross and backcross families. In addition, forewing bands were generally found to be unbroken. However, variation in the extent to which these bands showed no sign of the dumbbell

and belem spot was also notable, with some showing partial expression of these phenotypes. This variation was also evident between mapping families in the segregation ratios of broken band to unbroken band (Table 3.1). When using both the relaxed and stringent thresholds for the presence of the broken band (scoring methodology detailed in section 3.2.3), most families were found to show ratios that did not differ significantly from that expected for a recessive phenotype (broken) controlled by a single Mendelian locus (3:1 in an intercross; 1:1 in a backcross). However, others did differ significantly from these expected ratios, suggesting at least one other locus might be involved in the genetic control of this phenotype (see Table 3.1).

Family	Cross type	Broken	Unbroken	Sample size	$\chi 2$ (>2.5)	$\chi 2$ (=2, >1.5)
B5	F2	5 | 4	30 | 31	35	0.143	0.064
B8	BC	34 | 34	78 | 78	112	**<0.001**	**<0.001**
B10	F2	16 | 13	69 | 72	85	0.188	**0.039**
B11	F2	4 | 4	46 | 46	50	**0.006**	**0.006**
B12	F2	13 | 13	53 | 53	66	0.32	0.32
B13	F2	11 | 8	46 | 49	57	0.32	0.055
B14	BC	75 | 74	79 | 80	154	0.747	0.628

Table 3.1 – Variation in the ratios of broken to unbroken bands by mapping family. On the left of columns, bands are scored using the relaxed threshold of >=2.5 for presence of a broken band, on the right are bands scored using the stringent threshold for presence of a broken band (scoring methodology detailed in methods). In bold are $\chi 2$ p-values for the two scoring methods, showing significant deviation from the expected ratio for a recessive phenotype (broken) controlled by a single Mendelian locus.

Variation in the colour of red/orange pigmentation was noticeably greater on the ventral surface of the forewings relative to the dorsal sides across all families (demonstrated in B10 and B14 in Figure 3.4). Principal component analysis between *H. melpomene meriana* and *H. melpomene aglaope* identified one PC axis, PC1, which described approximately 91% of all variation and clearly separated the two subspecies (Figure 3.3). This was largely driven by variation in the amount of green recorded on both the ventral and dorsal

surfaces. The redder *H. melpomene meriana* phenotype appears to be partially dominant, with both F1s and the backcross progeny clustering closer to *H. melpomene meriana* than *H. melpomene aglaope*. However, these F1 phenotypes were not as extreme as those of *H. melpomene meriana,* with one of the six F1s tested skewed from the mean towards the *H. melpomene aglaope* phenotype, suggesting there may be more than one locus involved. F2 progeny exhibited more variation than back cross progeny, but with individuals skewed towards the redder *H. melpomene meriana* phenotype (Figure 3.3).

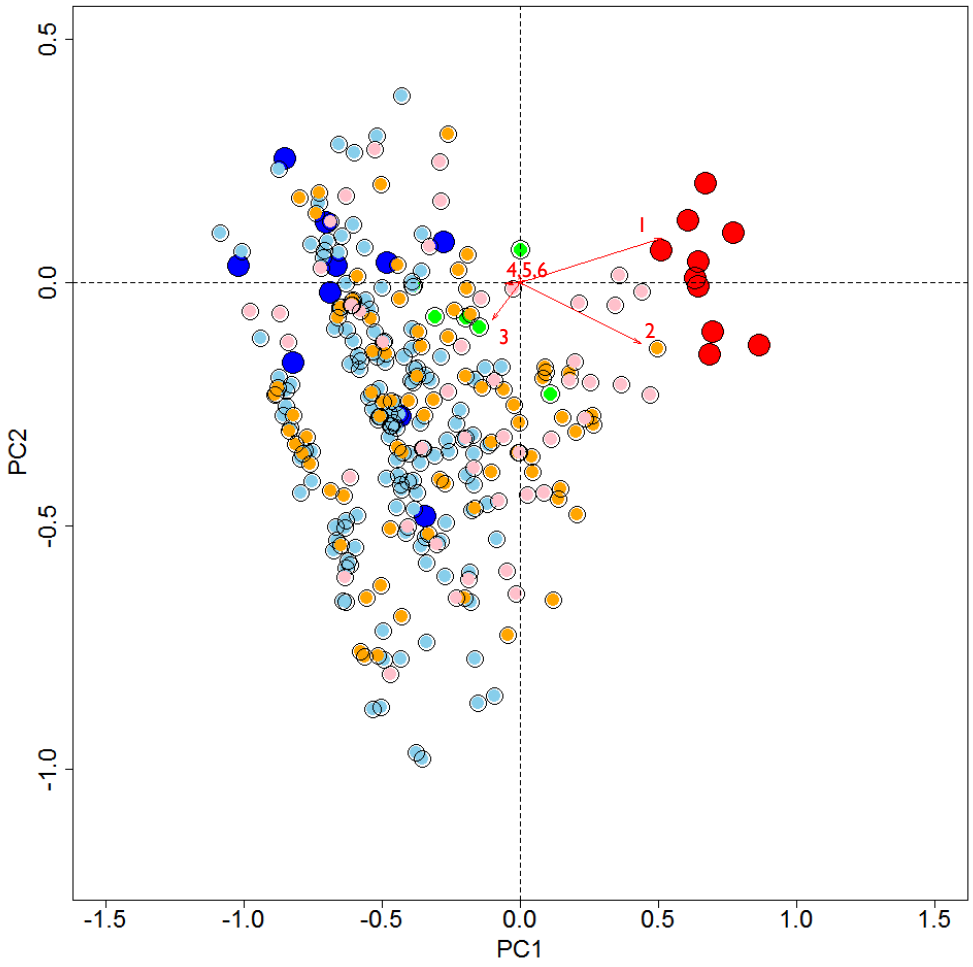


Figure 3.3– Principal component analysis of log10 transformed RGB values, from both ventral and dorsal sides. PC's describe variation between individuals of the two parental stocks. RGB values of F1s, F2s and back cross individuals were then transformed using eigenvectors onto these axes. Red – *H. melpomene aglaope*; dark blue – *H. melpomene meriana*; green – F1 samples; orange – F2 progeny from family B10; pink – F2 progeny from family B11; light blue – back cross progeny from B14. Variable loadings are shown as arrows: 1, dorsal green; 2 ventral green; 3, ventral blue; 4, ventral red; 5, dorsal blue and 6, dorsal red.

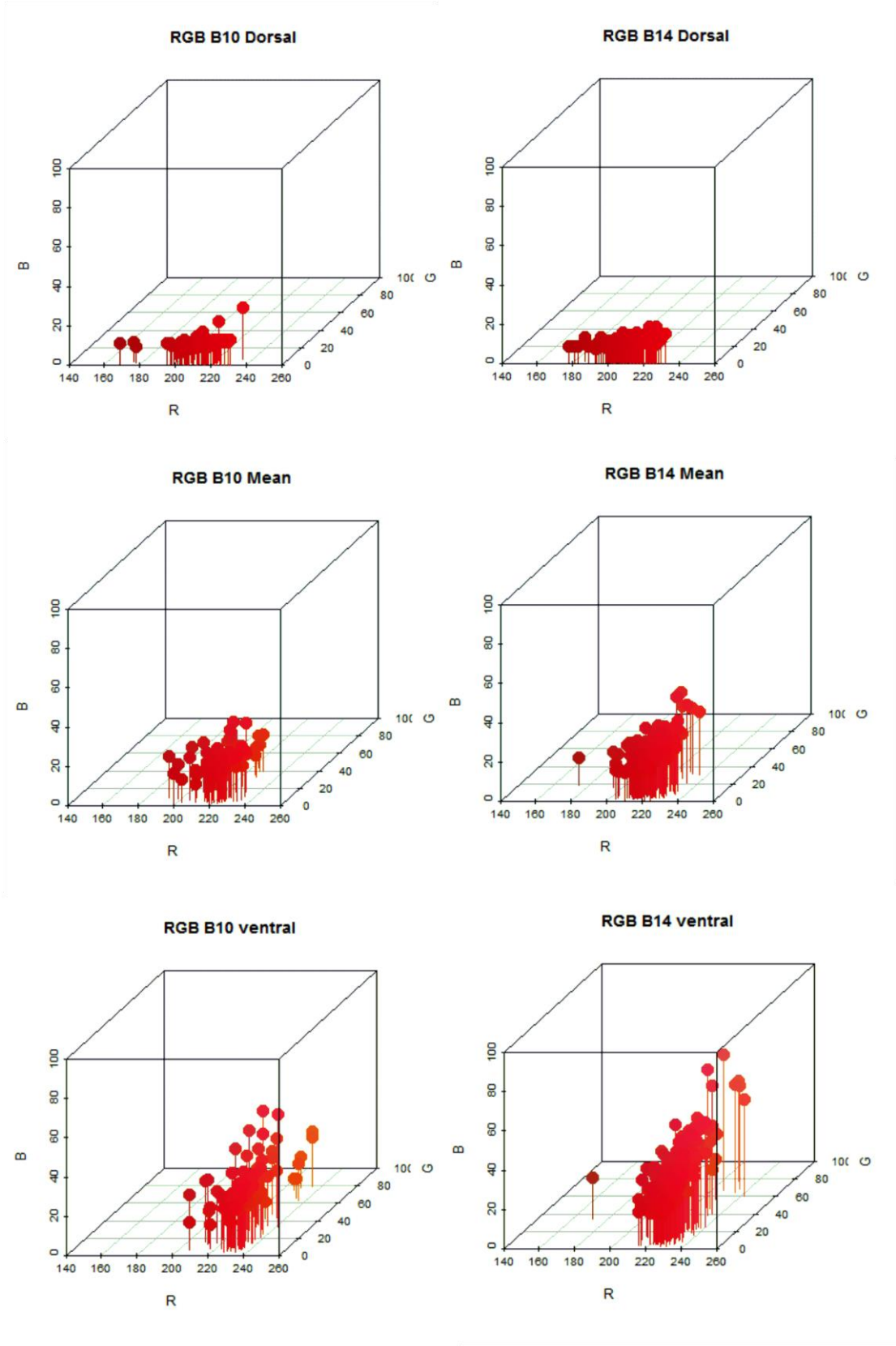


Figure 3.4 – Variation in red-orange pigmentation in families B10 and B14, plotted and coloured as RGB values. Top, dorsal variation; middle; mean variation; bottom ventral variation, in red orange colouration of the forewing dennis patch.

Medial band shape was variable within both the stocks of *H. melpomene meriana* and *H. melpomene aglaope*, with variation across most PC axes not showing distinct clusters between the two subspecies. However, across PC1, which explained 44% of the variation, the two subspecies did showed distinct but loose clustering (Figure 3.5). Across PC1, F1s had a somewhat intermediate phenotype, but that was still closer to that of *H. melpomene meriana*. In contrast, backcross progeny from family B14, were clearly more closely clustered with pure *H. melpomene meriana* individuals than those of *H. melpomene aglaope*. F2 progeny from B10 and B11 showed more variation than back cross samples, with some having more *H. melpomene meriana* phenotypes and others more *H. melpomene aglaope* phenotypes (see Figure 3.5a). Overall this pattern of variation suggests that there is at least one locus affecting medial band shape, and which also exhibits some dominance. However, the somewhat intermediate phenotypes of the F1s, combined with the variation along other PC axes, suggest at least one other locus plays a role in determining medial band shape (Figure 3.5b). The main effects of these PCs appear to be in changes to the shape of the last element of the medial band, as well as element two, and the distal edges of each element (Figure 3.6).

Based on these analyses of phenotypic variation, two mapping families were chosen to be genotyped for linkage map construction and QTL mapping analysis, these were the back cross family B14 and the F2 family B10. As well as good segregation of both the medial forewing band and pigmentation, B14 clearly followed the expected pattern of segregation for a recessive phenotype (broken) controlled by a single Mendelian locus for the broken band, while B10 appeared to vary from it, potentially meaning a second loci controlling this trait might be identified in this family. Finally, both families had large sample sizes and F1 fathers had been sampled, making the construction of the linkage map and pooling during QTL analysis substantially easier.

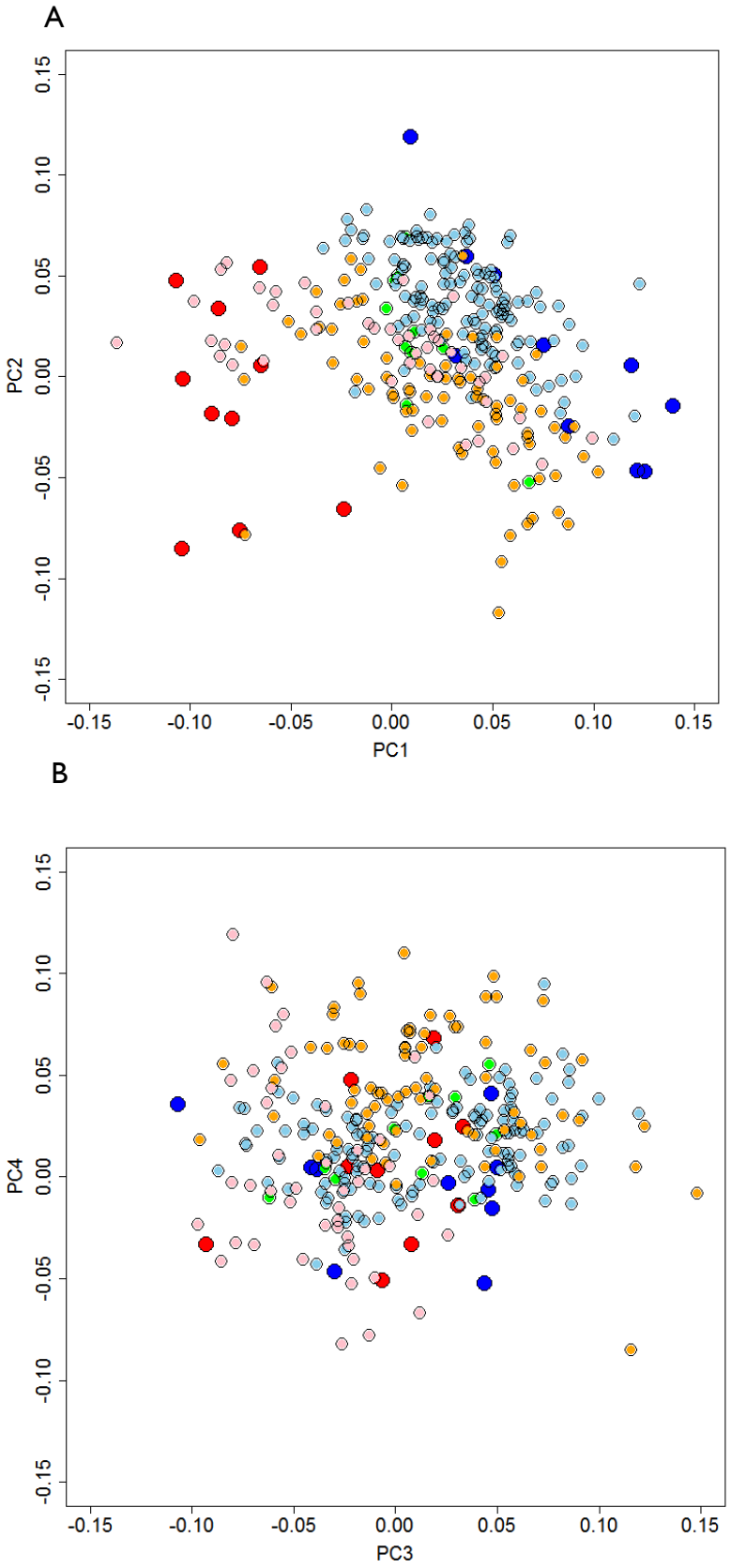


Figure 3.5 – Principal component analysis of medial band shape variation. PCs describe variation between individuals of the two parental stocks. F1s, F2s and back cross individuals were then transformed using eigenvectors onto these axes. Red – *H. melpomene aglaope*; dark blue – *H. melpomene meriana*; green – F1 samples; orange – F2 progeny from family B10; pink – F2 progeny from family B11; light blue – back cross progeny from B14. A) shows PCs 1 and 2; B) shows PCs 3 and 4.

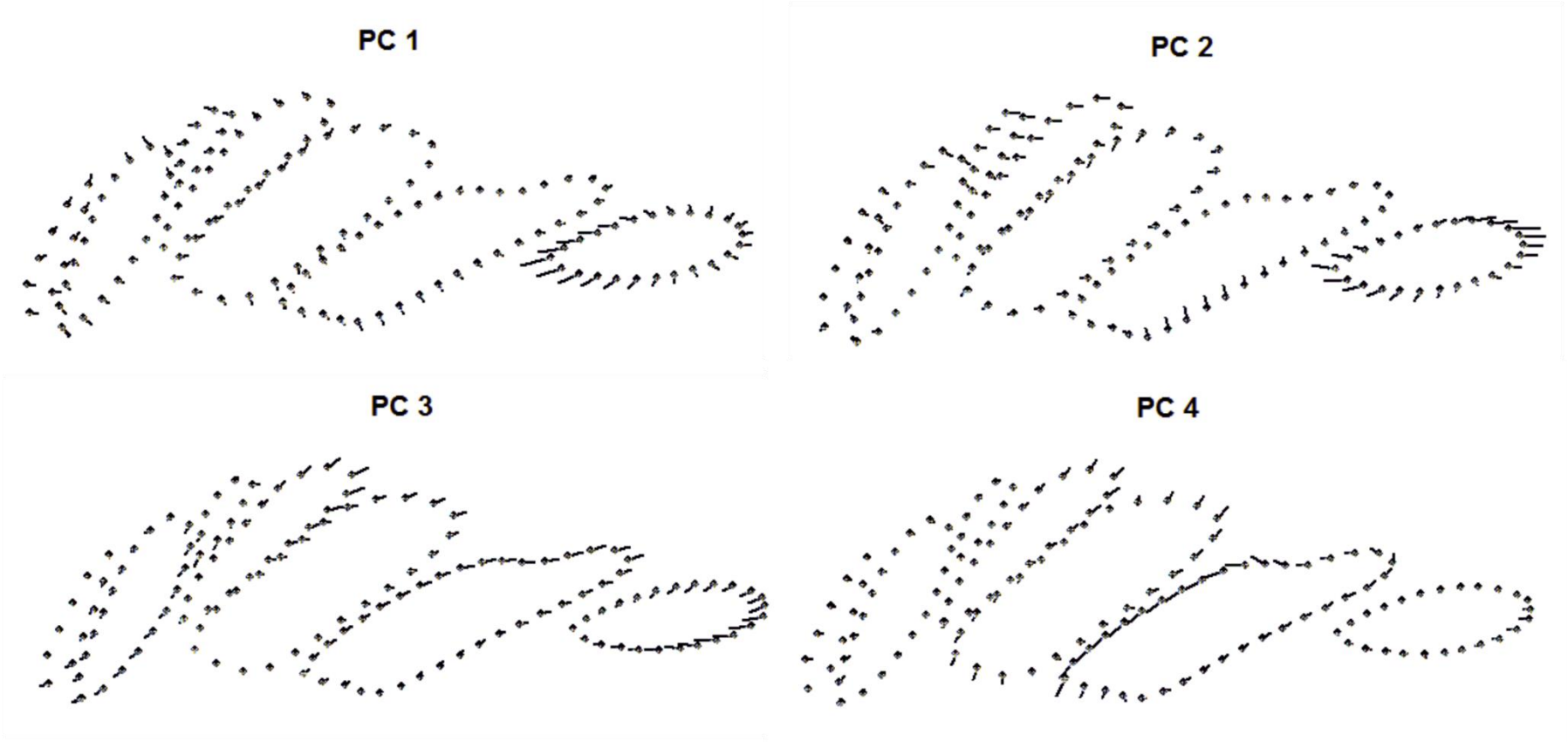


Figure 3.6 – The effects of each of the first four principal component axis, between stock individuals of *H. melpomene aglaope* and *H. melpomene meriana* on medial band shape. Principal component axis explain 44, 15, 12 and 6 percent of the total variation respectively. Elements are numbered from left to right, 1-7.

3.3.2 The linkage map

Before linkage map construction, one sample (PS360) was removed from Mapping family 14 due to very high levels of missing data, >95% compared to the second highest of 53% and an average of 9% across both mapping families. In addition, three samples (PS252, PS699 and PS703) were removed from B10, after showing a lower IBD than expected. This gave a final dataset of 219 progeny and four parents from which to construct maps of each linkage group from. Filtering data based on depth of coverage, genotype quality, missing data, and other parameters (detailed in the methods) gave a final set of ~150,000 good quality genotype markers from which to build a linkage map. Following the separation of linkage groups, approximately 26,000 paternally informative markers in B14, and 17,000 in B10 were assigned to linkage groups, while 54,305 markers remained as singular markers. After joining remaining singular markers to the linkage groups 46,804 paternally informative markers were found on linkage groups in B10 and 49,507 in B14. It should be noted that many of these markers were dual informative, especially in B10, due to the higher heterozygosity of the F1 mother of this mapping family relative to that of the mother of B14. In addition, many were only informative in one mapping family. The final linkage map was constructed from 3879 markers across 21 linkage groups composed of 1690.833 centiMorgans (see Table 3.2 and Figure 3.7a-d). While this linkage map is longer, it is still comparable to the known cM size (1,364.23 cM) of the *H. melpomene* genome (Davey *et al.* 2016). It should be noted that each marker names gives both scaffold and position of marker in the *H. melpomene* genome v2.

Linkage group	Number of Markers	Size cM
1	369	103.371
2	78	91.529
3	99	67.009
4	183	116.477
5	113	67.739
6	292	79.152
7	113	65.528
8	131	82.152
9	122	70.256
10	277	83.258
11	244	100.516
12	318	93.326
13	240	84.959
14	232	129.396
15	131	65.666
16	130	72.134
17	185	72.755
18	225	105.426
19	150	118.259
20	67	61.955
21(Z)	180	63.341
Total	**3879**	**1690.833**

Table 3.2 - The size in centiMorgans (cM) and number of markers across each linkage groups.

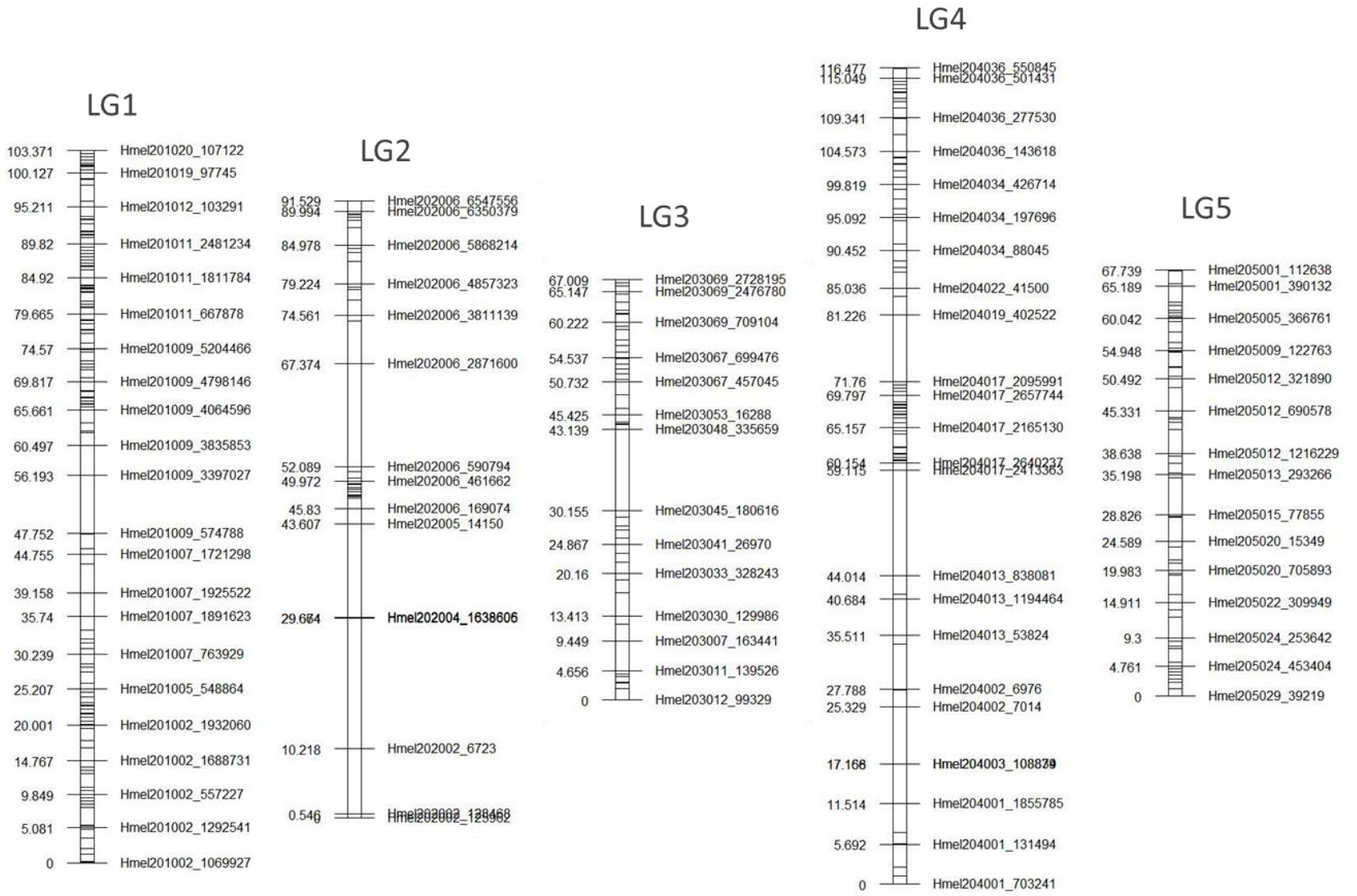


Figure 3.7a – Linkage groups one to five. Each line represents one marker. Only those markers closest to each 5 centiMorgan point are named, with their position given.

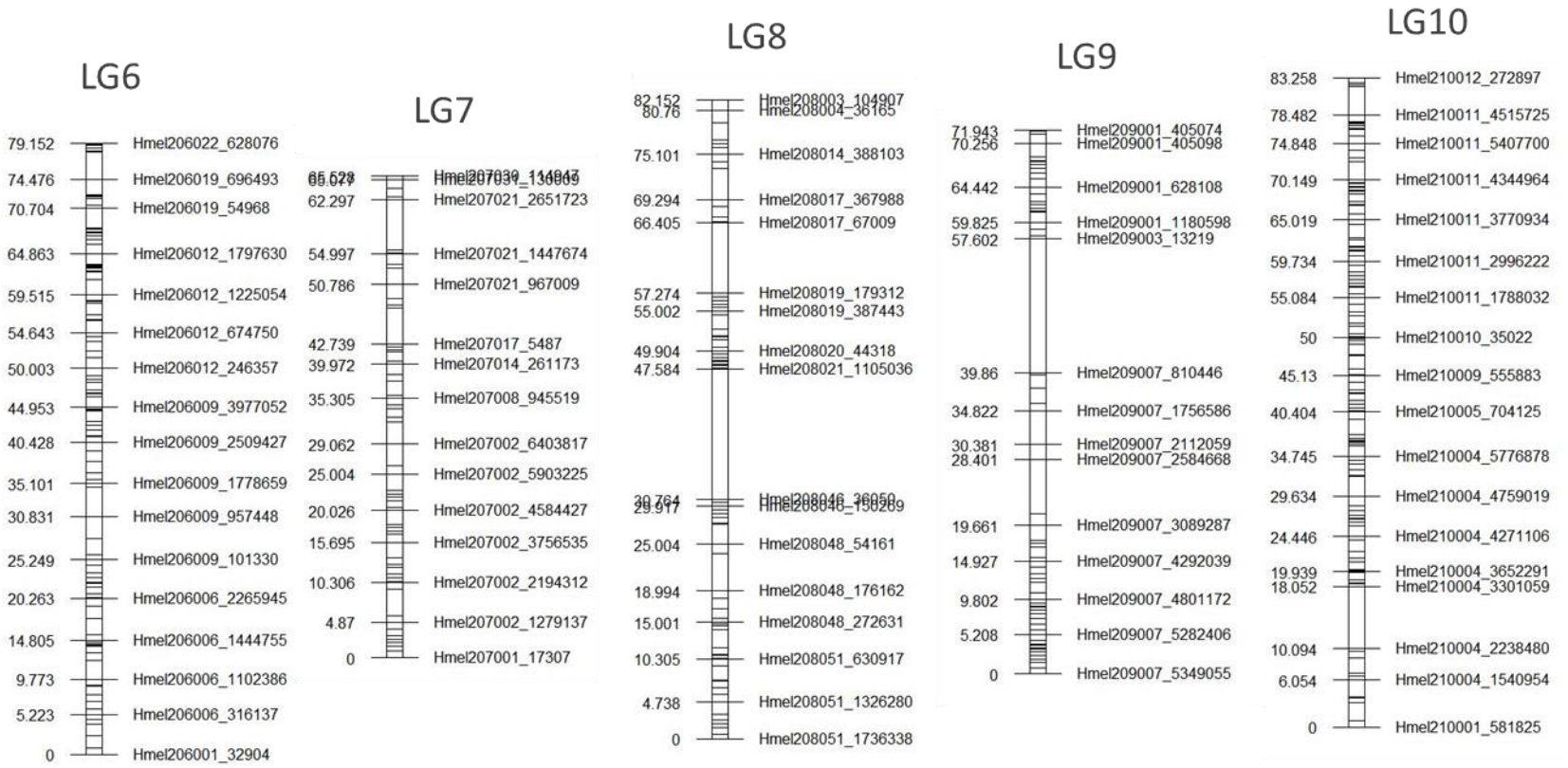


Figure 3.7b – Linkage groups six to ten. Each line represents one marker. Only those markers closest to each 5 centiMorgan point are named, with their position given.

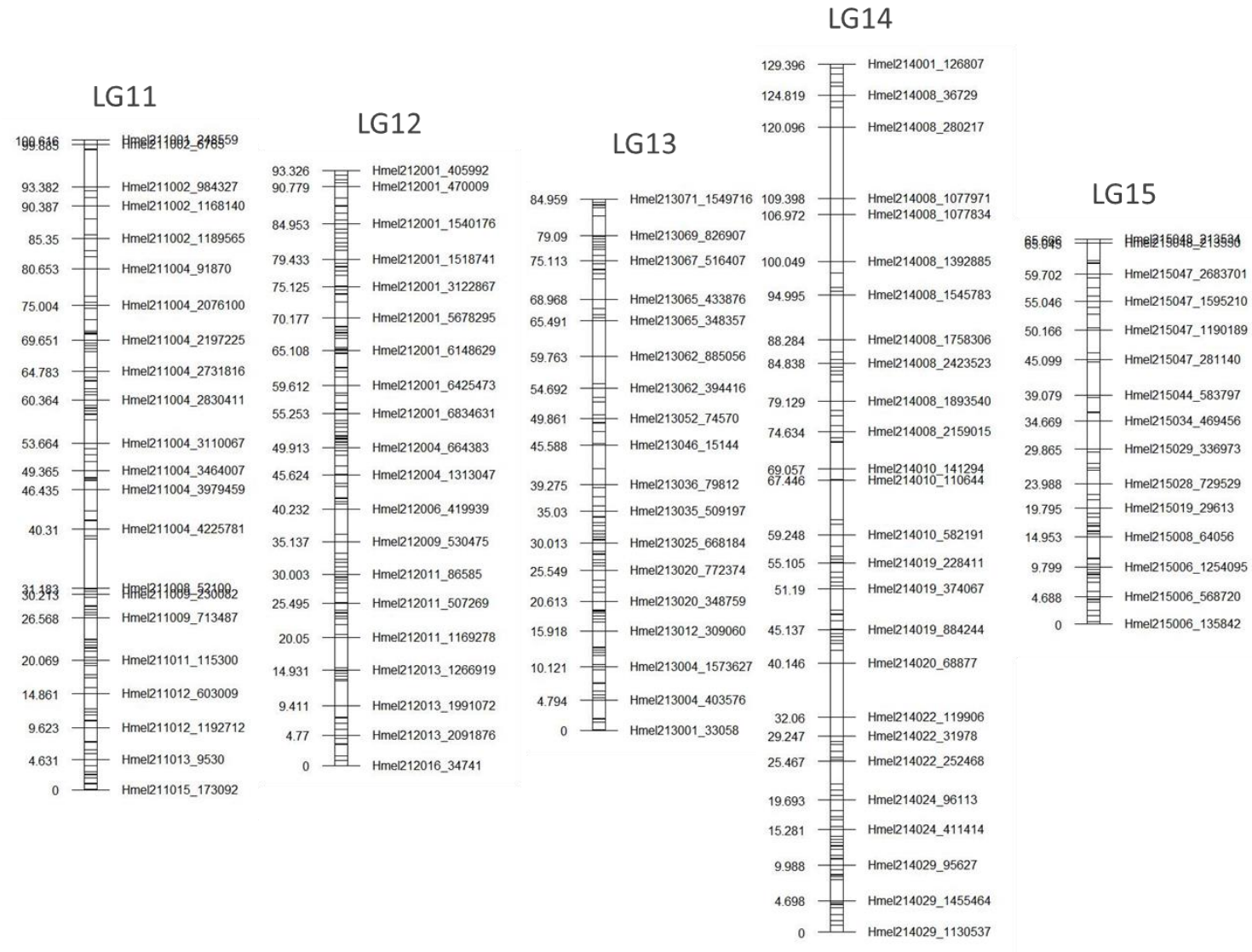


Figure 3.7c – Linkage groups eleven to fifteen. Each line represents one marker. Only those markers closest to each 5 centiMorgan point are named, with their position given.

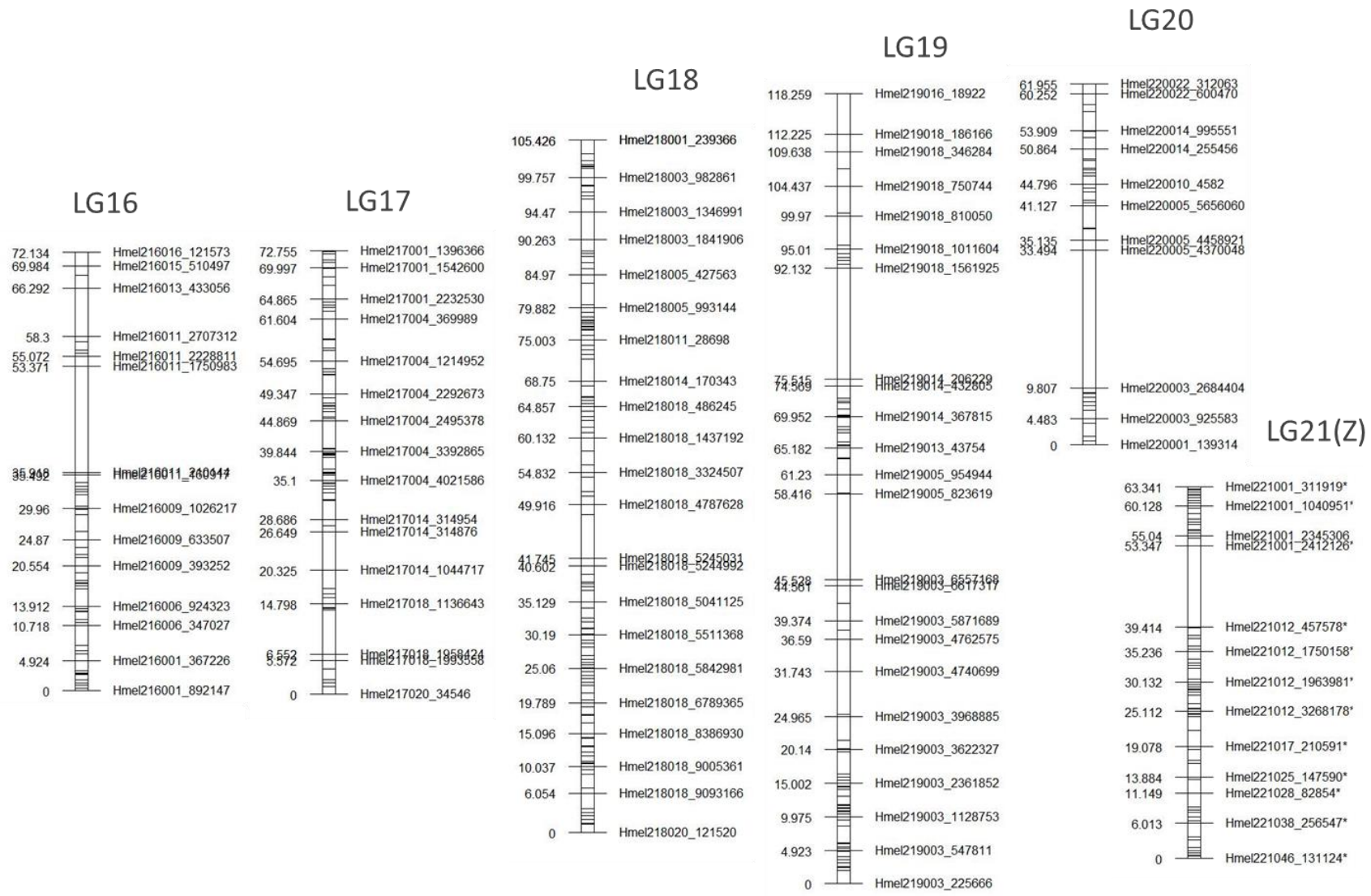


Figure 3.7d – Linkage groups sixteen to twenty-one. Each line represents one marker. Only those markers closest to each 5 centiMorgan point are named, with their position given.

3.3.3 **Genetic control of the rays phenotype**

82 individuals were phenotyped and genotyped from B10 and 136 from B14, giving a combined total of 218 progeny for QTL mapping analysis of rays. Only a single QTL on chromosome 18 was significant in the combined analyses on both mapping families (Figure 3.8) (LOD 55.39, $P < 0.004$), and in each of the analyses on the individual mapping families B10 (LOD 14.68, $P < 0.004$) and B14 (LOD 40.71, $P < 0.004$). Approximate 95% Bayesian confidence intervals from the combined analysis placed this QTL between 99cM and 100cM with LOD score highest at 99.74cM at the marker Hmel218003_990865. Intervals were identical from the analysis of backcross progeny from B14, with LOD score again highest at 99.74cM but at the marker Hmel218003_957111. In the F2 progeny from B10, LOD score was highest at 99.74cM at the marker Hmel218003_990865, with the QTL located within a wider region between 85.87cM and 100cM. These markers are the closest in the linkage map to the gene *optix* (705,604 - 706,407bp on scaffold Hmel218003) and to the known *rays* module (~800,000bp on scaffold Hmel218003) (Wallbank *et al.* 2016). Modelling the effects of these markers revealed that 63.07% of the overall variance was explained by the markers at this QTL.

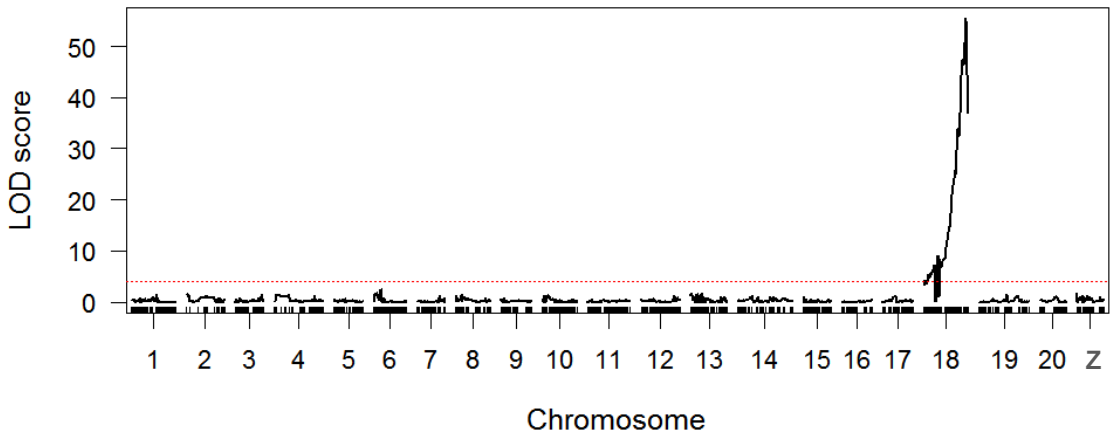


Figure 3.8 – Genome wide LOD scores from the combined family analysis of the rays phenotype, showing a highly significant major effect QTL on chromosome 18. Red line shows the genome wide threshold for significant LOD.

3.3.4 **Genetic control of the broken band**

In total 82 individuals were phenotyped and genotyped from B10, while 136 were from B14, giving a combined total of 218 progeny for QTL mapping of the broken band. In both the individual analyses of each mapping family and the combined analysis with both, LOD was greatest at a single locus on chromosome 10. In the combined analysis the LOD score at this locus was 49.8242 (LOD 49.82, $P < 0.004$) (Figure 3.10). Approximate 95% Bayesian confidence intervals placed this peak at a single marker 6.51cM. While using a LOD score drop-off of 1.5 placed the peak in a broader region between 6.05cM and 6.96cM, with the LOD score highest at the marker Hmel210004_1864446 (Figure 3.9). This marker is just 5kb away from the gene known as *WntA* (gene HMEL018100 of the *H. melpomene* v2 genome), supporting the role this gene is thought to play in controlling melanic patterning at the *Ac* region in *H. melpomene* (and at *sd* in *H. erato*) (Martin *et al.* 2012; Gallant *et al.* 2014a).

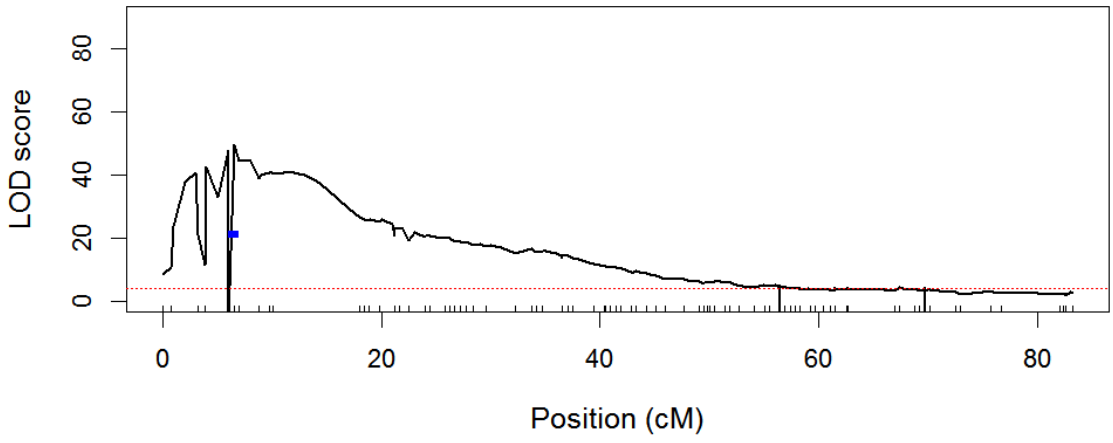


Figure 3.9 –LOD scores across chromosome 10 in the combined family analysis of the broken band phenotype. Red line shows the genome wide threshold for significant LOD, while the blue box indicates the LOD interval calculated as a drop of 1.5. Markers are identified as tick marks below.

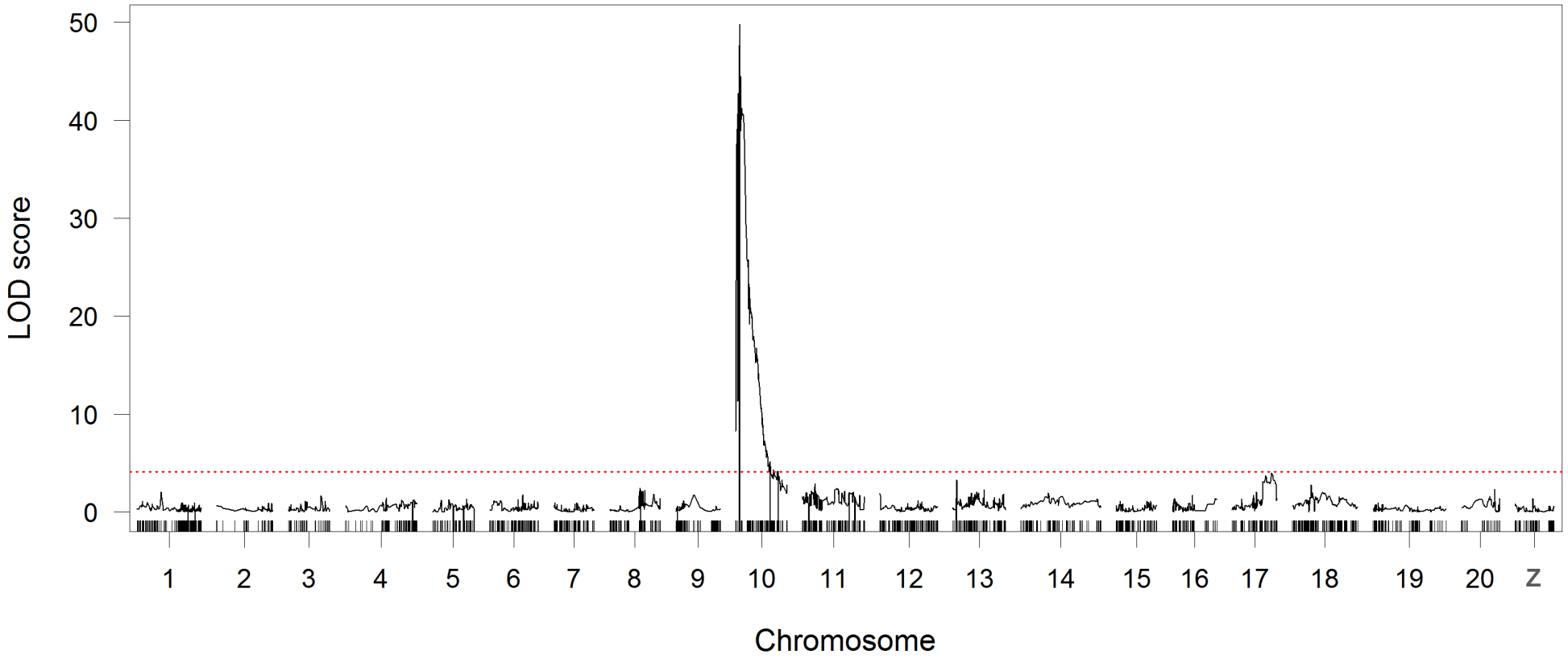


Figure 3.10– Genome wide LOD scores from the combined family analysis of the broken band phenotype, showing a highly significant major effect QTL on chromosome 10, and a putative QTL significant at the 0.1 P-value threshold seen only in mapping family B10. Red line shows the genome wide threshold for significant LOD.

In addition to this significant QTL on chromosome 10, there was an additional peak of LOD 3.98, on chromosome 17. In the combined analysis this peak was not significant at the *P = 0.05* threshold, but it was significant at the *P = 0.1* threshold (*P = 0.069*) Figure 3.10). Bayesian 95% confidence intervals placed this putative QTL within a broad region between 50cM and 69cM, with the highest LOD score at 63.92cM, at the marker Hmel217001_2592898. In all models, using both the combined family data and each mapping family separately, the percentage of the variance explained by the locus on chromosome 10 was always large (Table 3.3). In contrast, the locus on chromosome 17 explained only 1% of the variation in the additive model fit to the combined data with mapping family as a covariate.

Analysis	QTL1 only	QTL2 only	QTL1 with 2	QTL2 with 1
Combined + Fam	60.05*	7.43*	**53.7***	**1.08***
B14 (BC)	**74.98***	0.87	74.11*	0
B10 (F2)	39.36*	19.90*	**24.72***	**5.25***

Table 3.3 – Table showing the percentage of the overall phenotypic variance explained by each broken band locus in the various models fitted. With combined analysis and each family sperately. QTL1 and QTL2 are respectively the chromosome 10 and chromosome 17 loci. * indicates that the chi-squared P-value was significant (P < 0.05), these should be treated with caution as they are pointwise and so do not account for the search over the whole genome. In bold are models that had highest LOD fit.

The difference in the modelled effects of this locus, in the individual families, was also striking. The locus on chromosome 17, explained none of the overall variation in the backcross progeny of B14, when included in a model with the chromosome 10 loci. However, in the F2 progeny this chromosome 17 locus explained 5.25% of the overall variation with the chromosome 10 loci, and 19.9% when considered alone (Table 3.3). This difference in the effects between the two families can also be seen in the genome wide LOD scores for each mapping family (Figure 3.11). It is possible this chromosome 17 locus might at least in part explain the variable segregation patterns of this phenotype

seen between the various different mapping families examined for the analysis of phenotypic segregation.

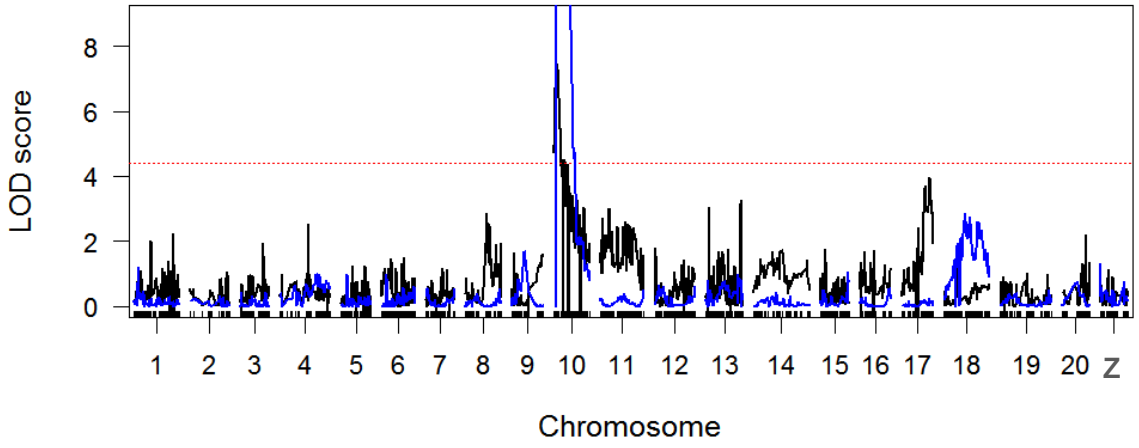


Figure 3.11- Genome wide LOD scores (cut off at LOD = 9), for the individual family analyses of the broken band phenotype (in B14 chromosome 10 peak goes up to LOD 41). Blue, B14 family; black, B10 family. The red line shows the genome wide threshold for significant LOD from the combined analyses, for separate analyses these thresholds are greater.

3.3.5 Genetic control of red-orange pigmentation

A total of 80 individuals from mapping family B10 were successfully phenotyped, while 136 were successfully phenotyped from mapping family B14. This gave a combined total of 216 progeny across the two families for use in QTL mapping analyses of red-orange pigmentation. Principal component analysis was first carried out on $\log_{10}$ transformed RGB values, across individuals from both mapping families. This analysis was concordant with that of the parental RGB colour values, as it was clear that variation in colour was dominated by variation in green and blue scores (Figure 3.12), with the largest variation driven by ventral green values making up most of PC1. This axis described 52% of the overall variation, while PC2, PC3 and PC4 respectively each described 33, 12 and 3 percent of the overall variation.

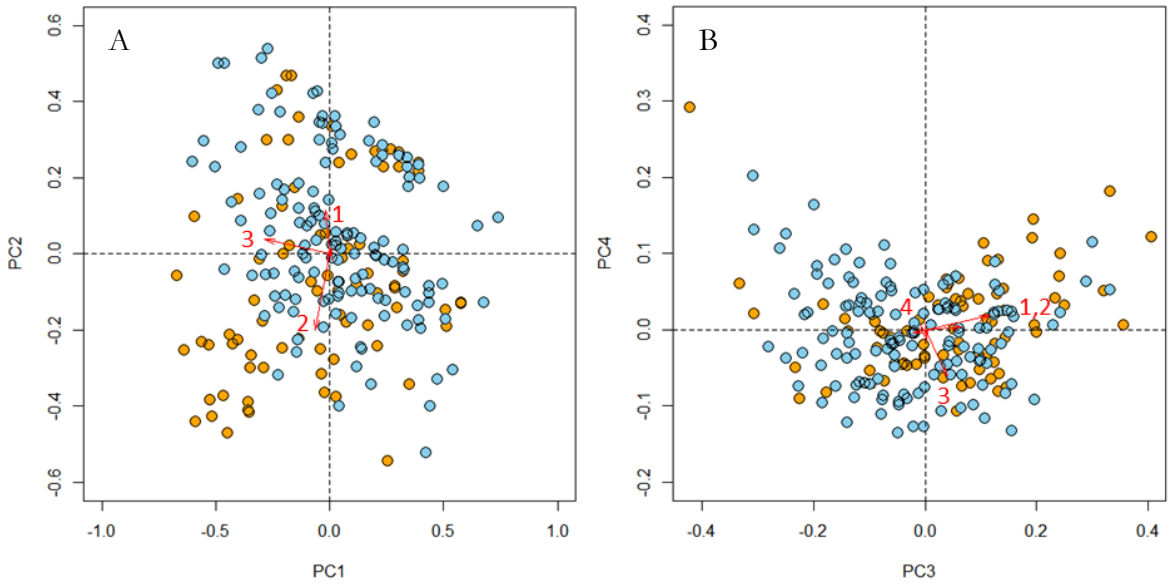


Figure 3.12 – Principal component analysis of log_{10} transformed RGB values, from both ventral and dorsal sides. PC's describe variation between individuals in the two mapping families. Orange – F2 progeny from family B10; light blue – back cross progeny from B14. Variable loadings showing sizeable variation are shown as red arrows. A) 1, ventral blue; 2, dorsal green and 3, ventral green. B) 1, ventral blue; 2, dorsal green; 3, dorsal blue and 4, ventral green.

In addition to analysing these two families together, with family as an additive covariate, each was also analysed separately, in order to confirm that these peaks were seen consistently across the two mapping families (Figure 3.15). A peak could be clearly seen in both mapping families on chromosome 13. However, this QTL was only significant in the F2 family B10 (B10, LOD 7.14, *P < 0.004*; B14, LOD 3.13, *P = 0.095*), likely due to the high ratio of noise to signal in phenotyping the backcross progeny from B14 due to the reduced phenotypic variation in the backcross progeny in comparison to that of the F2 progeny. In the F2 progeny Bayesian 95% confidence intervals placed this QTL in a wide region between 41.79cM and 61cM. Refining the positions of this QTL placed the locus on chromosome 13 closest to the marker Hmel213049_709945 at 49.15cM. Peaks could also be seen on chromosome 15 in both mapping families, though these were not significant and did not overlap entirely. In addition, a significant QTL was revealed on chromosome 18 in the backcross family B14 (LOD 4.2, *P = 0.015*). Bayesian 95% confidence intervals placed this QTL in a wide region between 90cM and 104cM, with

LOD highest at 101.15cM at the marker Hmel218003_325262. This marker is on the same scaffold as the gene *optix* which is known to be involved in red-orange element patterning, and which is located ~375kb away (Reed *et al.* 2011; Martin *et al.* 2014b; Wallbank *et al.* 2016).

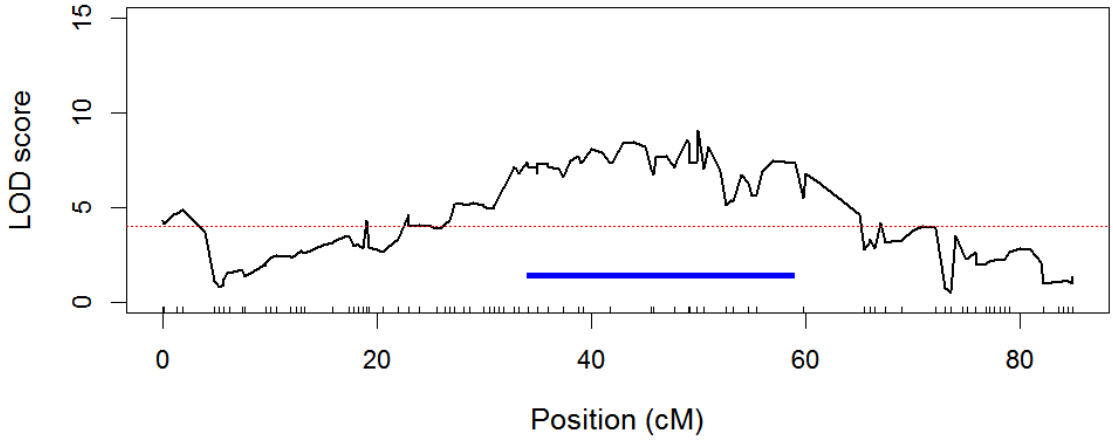


Figure 3.13 – LOD scores across chromosome 13 in the combined family analysis on log_{10} transformed ventral green RGB scores. Red line shows the genome wide threshold for significant LOD, while the blue box indicates the Bayesian 95% confidence intervals for this QTL. Markers are identified as tick marks below.

As fifty percent of the overall variation was described by variation in RGB ventral green values alone, this univariate phenotype was first used alone for QTL analysis. Genome wide LOD scores using the combined data from both mapping families, identified two significant QTLs (Figure 3.14), the largest being on chromosome 13, with a LOD score of 9.5 ($P < 0.002$), and the smaller being on chromosome 15 with a LOD score of 4.09 ($P = 0.046$). Bayesian 95% confidence intervals for both peaks were wide, with the peak on chromosome 13 between 34.0cM and 59.0cM, and the peak on chromosome 15 between 0cM and 50.63cM. LOD intervals with a drop of 1.5, were slightly narrower for chromosome 13, placing it between 38.0cM and 52.0cM (Figure 3.13). Refining the

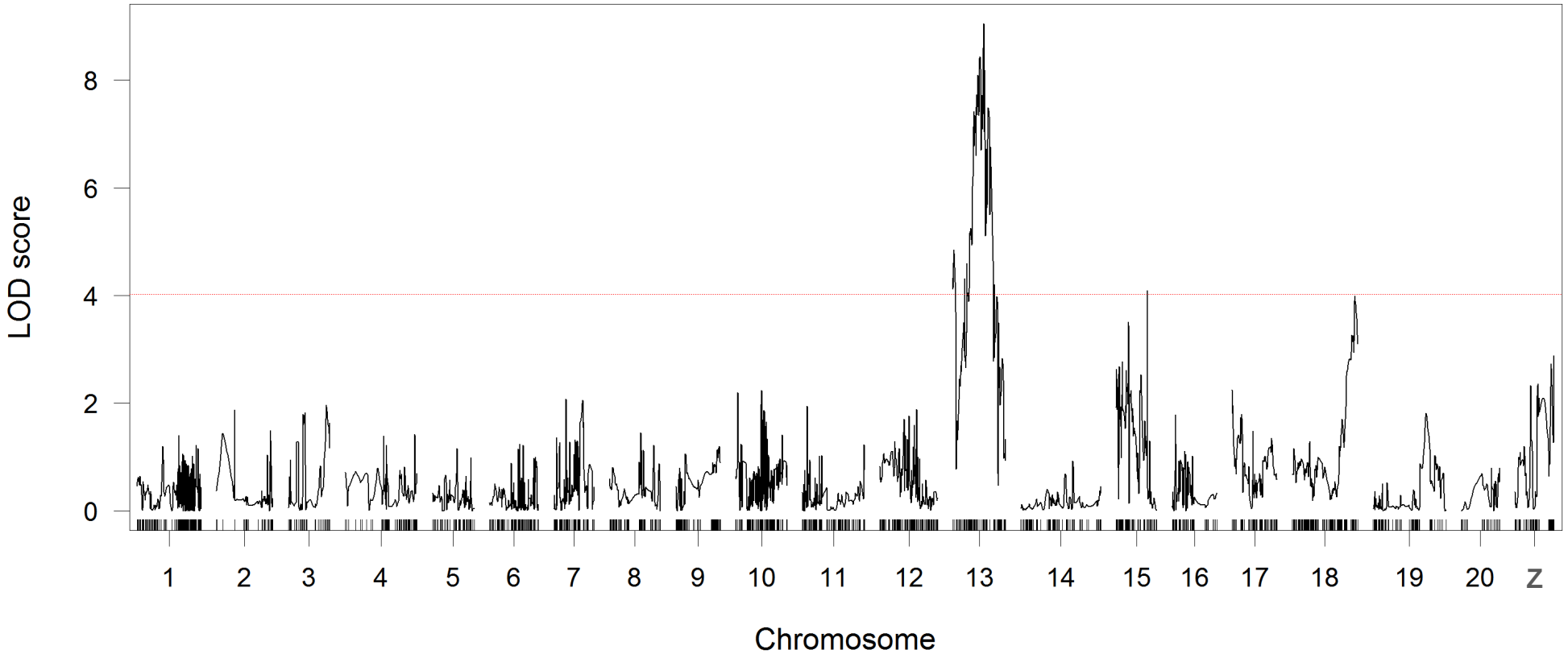


Figure 3.14 – Genome wide LOD scores from the univariate combined family analysis on log10 transformed ventral green RGB scores, showing significant QTLs on chromosomes 13 and 15. The red line shows the genome wide threshold for significant LOD.

positions of the QTLs with refineqtl(), placed the locus on chromosome 13 closest to the marker Hmel213051_109947 at 49.15cM, while the locus on chromosome 15 was placed closest to the marker Hmel215006_1599915 at 10.23cM. Modelling the effects of these two refined loci showed that the chromosome 10 locus explained a much greater percent of the overall variation, 17.5%, while the chromosome 15 locus explained 4.4%.

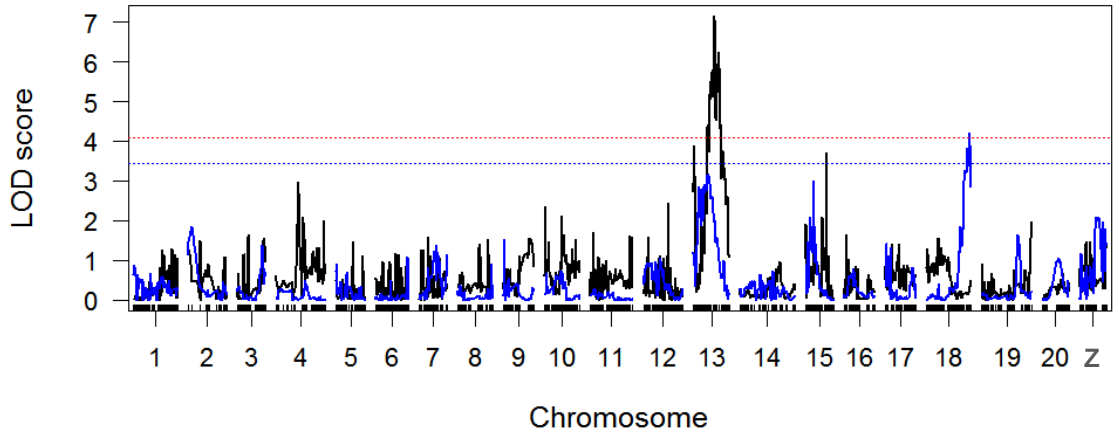


Figure 3.15 - Genome wide LOD scores, for the univariate individual family analyses, on log10 transformed ventral green RGB scores. Blue, B14 family; black, B10 family. The red line shows the genome wide threshold for significant LOD from the combined analyses, for separate analyses these thresholds are greater.

As already discussed, the principal component analyses on all transformed RGB values from both families, the first four principal components were found to explain 52%, 33 %, 12% and 3% of the overall variation respectively, these were all included in a multivariate QTL mapping analysis. LOD scores from across the genome, using the combined data from both mapping families, again identified the main QTL as being on chromosome 13 and as being highly significant (LOD 19.21, *P = 0.004*). Under this additive model, a QTL on chromosome 15 was also again found to be significant (LOD 4.37, *P = 0.027*), this is congruent the results from the univariate ventral green analyses (Figure 3.17).

Bayesian 95% confidence intervals placed the location of the QTL on chromosome 13 between 49cM and 60cM, with LOD score highest at the marker Hmel213051_54727 at 49.14515cM (Figure 3.16); this is a narrower window than that found in the univariate

ventral green analysis. This peak is very close to the refined position on chromosome 13 from the univariate log10 transformed ventral green analysis, which was at the marker Hmel213051 _109947 at 49.14716cM. Further, while the univariate analysis had also placed the peak on chromosome 15 somewhere within a very wide region between 0cM and 50.63cM, the Bayesian 95% confidence intervals for this peak in the multivariate analyses were narrower, between 0cM and 33cM, and centred on 17cM.

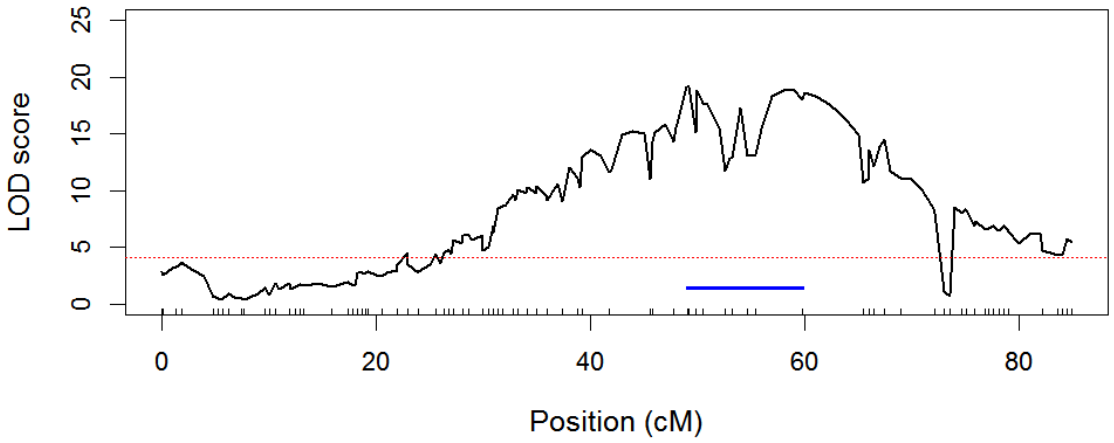


Figure 3.16 – LOD scores across chromosome 13 from the multivariate, combined family analysis on log10 transformed RGB scores. Red line shows the genome wide threshold for significant LOD, while the blue box indicates the Bayesian 95% confidence intervals for this QTL. Markers are identified as tick marks below.

As well as this combined analyses, genome wide LOD scores were also calculated separately for each of the individual mapping families (Figure 3.18). In both families the locus on chromosome 13 was identified as significant (LOD 9.59, $P = 0.012$ for the F2 family B10; LOD 10.28, $P < 0.004$ for the backcross family B14), and as having the highest LOD score across the genome. No other significant peaks were found in the backcross progeny from B14, although LOD was raised on chromsome 18, being very close to significance (LOD 3.97, $P = 0.051$). A significant QTL was though, identified on chromosome 17 (LOD 5.89, $P = 0.048$), in the F2 family, B10. Again these marginal P-values should be treated with some degree of caution.

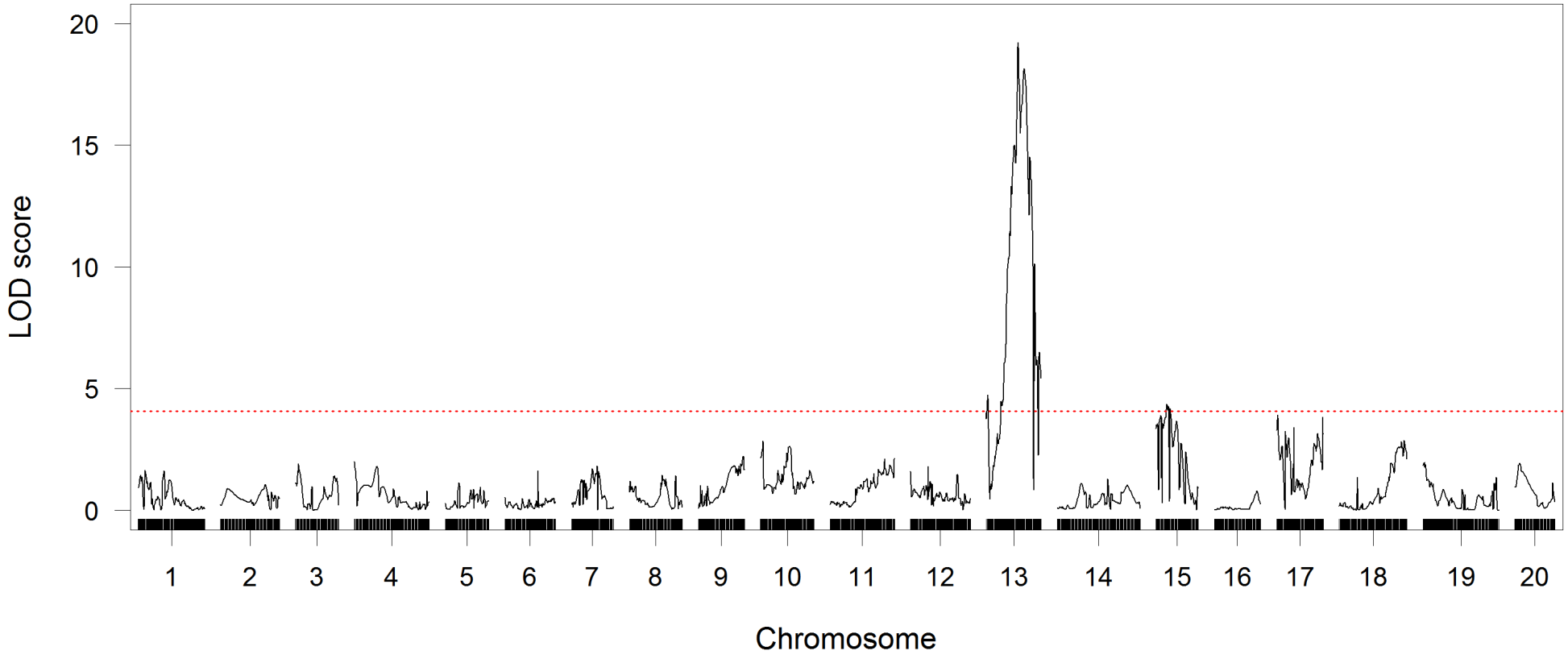


Figure 3.17 – Genome wide LOD scores under an additive model, from the combined family, multivariate analysis on log10 transformed RGB scores, showing significant QTLs on chromosome 13 and 15. Red dashed line shows genome wide threshold for significant LOD. Note: Z chromosome could not be included for QTL scans of multivariate traits.

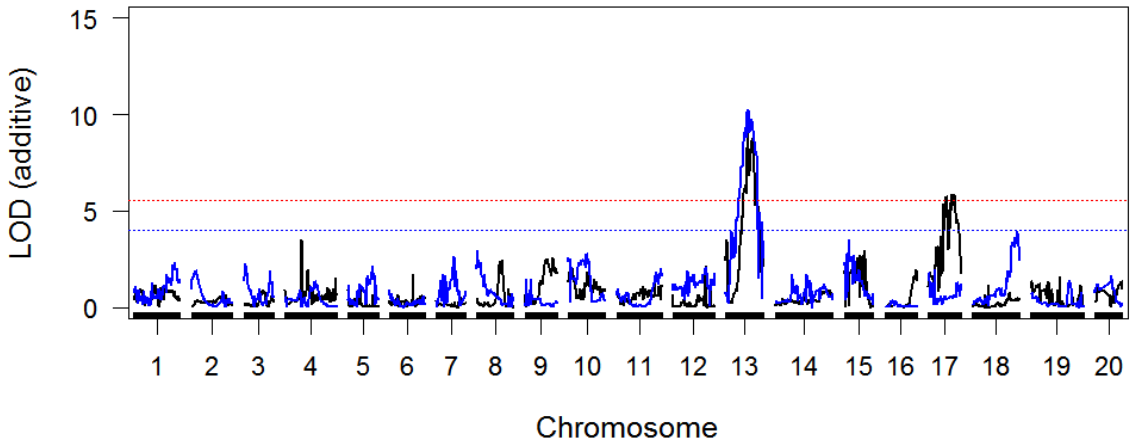


Figure 3.18 – Genome wide LOD scores, for the multivariate individual family analyses, on all log10 transformed RGB scores. Blue, B14 backcross family; black, B10 F2 family. Red line shows the genome wide threshold for significant LOD for the B10 backcross progeny, and the blue line shows the genome wide threshold for significant LOD for the B14 F2 progeny.

In the backcross progeny approximate 95% Bayesian confidence intervals positioned the QTL on chromosome 13 between 42cM and 62cM with LOD highest at 49.84cM at the marker Hmel213052_119960, while in the F2 progeny from B10 this QTL was positioned between 48cM and 62cM with LOD highest at 49.15cM at the marker Hmel213051_96232 in B10. These positions are broadly consistent with those found in the univariate and multivariate combined family analyses of red-orange pigmentation. For B14, Bayesian 95% confidence intervals placed the almost significant QTL on chromosome 18 within a region consistent with that found in the univariate analyses, on this same family, with the QTL within a slightly wider region between 88cM and 104cM, but with LOD highest at the same marker Hmel218003_325262. Bayesian 95% confidence intervals also placed the significant QTL found on chromosome 17 in the F2 mapping family B10, within a wider region between 34.64cM and 63.46cM, with LOD highest at 58.28cM at the marker Hmel217004_785124. This region overlaps with the location of the putative QTL found to possibly contribute to the genetic control of the broken band in this mapping family, with the Bayesian confidence intervals for that broken band QTL between 50cM and 69cM, with LOD highest at 63.92cM at the marker Hmel217001 _2592898.

3.3.6 **Genetic control of medial band shape**

A total of 79 individuals from mapping family B10 were successfully phenotyped, while 134 were successfully phenotyped from mapping family B14 for medial band shape. This gave a combined total of 213 progeny across the two families for use in the combined analyses. Generalised Procrustes analysis and principal component analyses was carried out on this combined dataset, so that the all variation among progeny could be mapped. All Principal components explaining over 1% of the variation were used for the QTL mapping analysis. In the combined analysis with both the F2 progeny from B10 and the backcross progeny from B14 these explained over 88.7% of the total variation in medial band shape. A genome wide QTL scan on these PC axes revealed a number of significant QTLs under the additive model (Figure 3.19). Strikingly the highest LOD score was again on chromosome 10 at the same position 6.51cM (LOD 39.73, $P < 0.004$) as the main locus found to be controlling the broken band. Additional loci were also found to be significant on chromosomes 9 (LOD 9.23, $P = 0.001$), 12 (LOD 4.57, $P = 0.035$), 13 (LOD 10.26, $P = 0.001$), 15 (LOD 8.95, $P = 0.002$), 17 (LOD 5.38, $P = 0.017$), 18 (LOD 11.38, $P < 0.004$) and 20 (LOD 5.88, $P = 0.011$).

95% Bayesian confidence intervals placed this main peak on chromosome 10, between 6.51cM and 7cM, with LOD highest at 6.51cM, at marker Hmel210004_1753431. This marker was one of two identified as being within the 95% Bayesian confidence intervals for the QTL on chromosome 10 controlling the presence or absence of the broken band. Two more of these QTLs also appeared to overlap with QTLs identified in other analyses, those on chromosome 17 and 13. The 95% Bayesian confidence intervals for the QTL on chromosome 13 placed this between 35.96cM and 52cM with LOD highest 64.85cM. This overlaps with the position of the QTL identified as playing a role in orange-red pigmentation which was located within a window between 49cM and 60cM on

chromosome 13. In addition, LOD scores within this region were greatest at markers very close to one another, at the marker Hmel213049_709945 at 49.147cM for medial band shape and at the marker Hmel213051_54727 at 49.145cM for red-orange pigmentation.

Bayesian 95% confidence intervals for the chromosome 17 locus placed this QTL within a region between 53.42cM and 71.35cM, with LOD highest at the marker Hmel217001_2232440 at 64.85cM (Table 3.4). Interestingly this is the region identified in the F2 progeny from B10 that was putatively identified as being involved in the control of the broken band and in orange-red pigmentation. For the broken band, 95% Bayesian confidence intervals had placed this putative QTL was within a broad region between 50cM and 69cM, with LOD highest at 63.92cM. For orange-red pigmentation the significant chromosome 17 locus in the F2 mapping family B1 had been placed within a region between 34.64cM and 63.46cM, with LOD highest at 58.28cM at the marker Hmel217004_785124. The 95% Bayesian intervals for other QTLs located on chromosomes 9, 12, 15, 18 and 20 are detailed in Table 3.4.

Chromosome	9	10	12	13	15	17	18	20
LOD	9.23	39.73	5.62	10.25	8.95	5.36	11.38	5.88
Lower (cM)	48	6.51	18	35.96	36	53.42	10	5.96
Highest (cM)	53	6.51	28.64	49.15	36.156	64.85	19.78	9.8
Upper (cM)	60	7	91	52	36.156	71.35	58.05	16

Table 3.4 – Bayesian 95% confidence intervals of QTLs identified from the combined analysis of medial forewing band shape

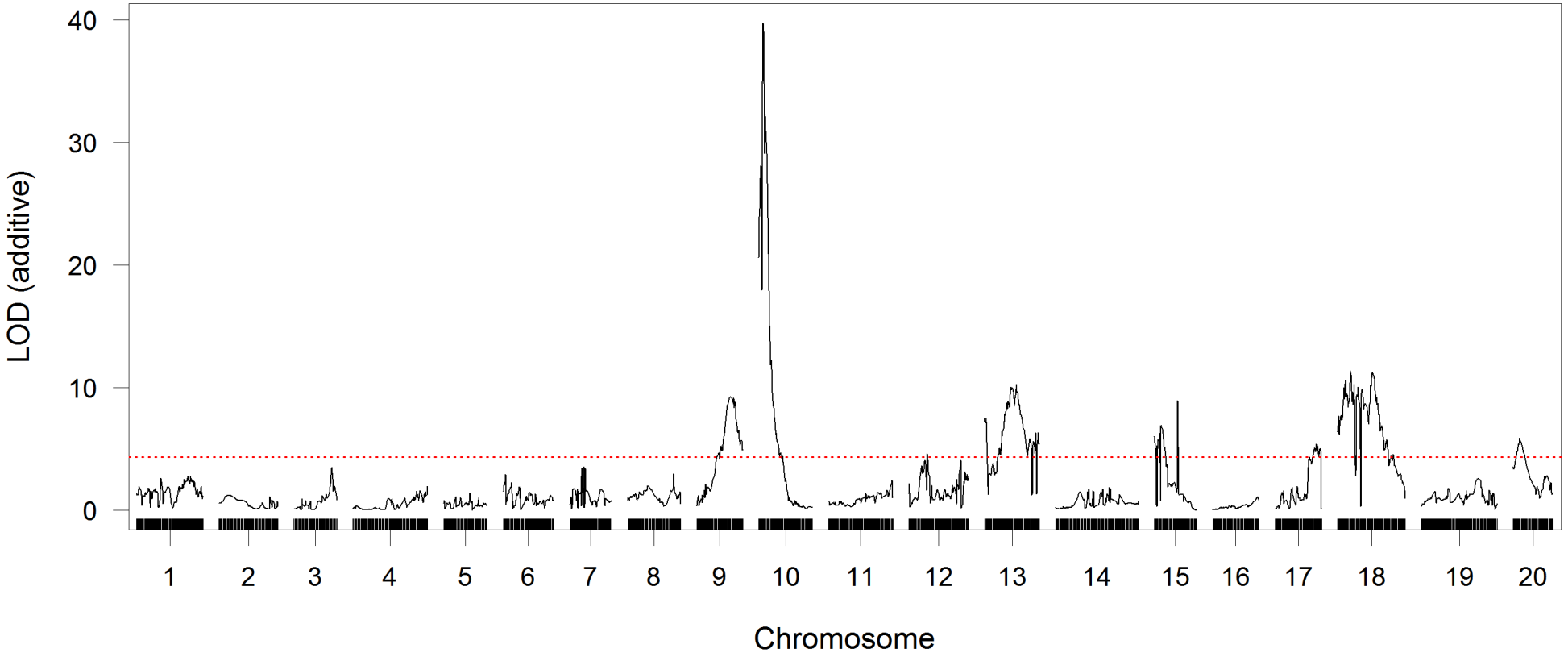


Figure 3.19 – Genome wide LOD scores, for the combined analysis of medial band shape under a strictly additive model. Significant QTLs are found on chromosomes 9, 10, 13, 15, 17 and 18. Red line shows the genome wide threshold for significant LOD. Note: Z chromosome could not be included for QTL scans of multivariate traits.

Genome scans on each of the separate mapping families revealed that QTLs were only consistently found in both, on three chromosomes (Figure 3.20); chromosomes 15 (in B14: LOD 7.86, P < 0.002; in B10: LOD 5.33, P < 0.045), 18 (in B14: LOD 6.65, P = 0.004; in B10: LOD 7.31, P < 0.001) and 10 (in B14: LOD 43.48, P < 0.004; in B10: LOD 6.57, P < 0.019). In addition to these QTLs, a number of other significant QTLs were identified in the backcross progeny from B14. These were the QTLs previously identified from the overall analyses; on chromosomes 9 (LOD 5.80, P = 0.007), 13 (LOD 13.84, P < 0.004), and 20 (LOD 5.59, P < 0.010). In the F2 progeny, no other significant QTLs were identified at the *P = 0.05* threshold, with the QTL on chromosome 17 from the combined analysis not significant (P = 0.065). While there was some evidence of raised LOD scores in the F2 progeny, around the QTL on chromosome 9, there was no rise in the LOD scores in the F2 progeny around the other loci seen in the backcross progeny. In addition, the QTL previously identified on chromosome 12 in the combined analyses was found to not be significant in either of the two mapping families.

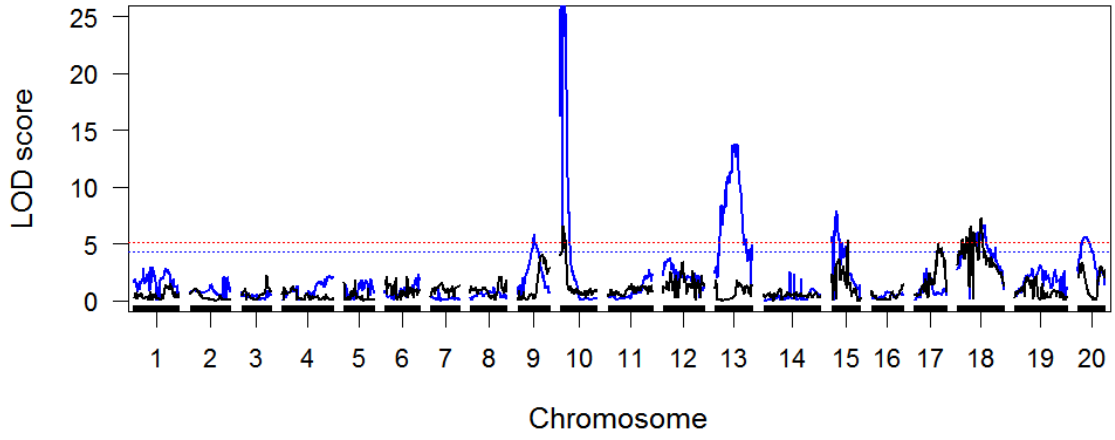


Figure 3.20– Genome wide LOD scores, for the individual family analyses of medial band shape. Blue, B14 backcross family; black, B10 F2 family. Red line shows the genome wide threshold for significant LOD for the B10 backcross progeny, and the blue line shows the genome wide threshold for significant LOD for the B14 F2 progeny.

The 95% Bayesian intervals were fairly narrow in both analyses for the QTL on chromosome 10 (Table 3.5), and as expected corresponded tightly to the region identified in the combined analyses. Likewise the locations of greatest LOD for the QTL on chromosome 18 were very close to each other, at 54.81cM and 57cM, as were the upper and lower intervals (Table 3.5), this suggests that these QTLs are one and the same. As expected, the approximate 95% Bayesian confidence intervals for the QTL on chromosome 13, in mapping family B14, corresponded tightly to the intervals for this QTL in the combined analysis. In addition, the 95% Bayesian intervals around the highest LOD points on the other chromosomes with putative QTLs in the combined analysis were generally found to overlap with those from the combined analyses, these intervals are detailed in Table 3.4 and Table 3.5.

Chromosome	9	10	12	13	15	17	18	20
LOD	**5.80**	**43.48**	3.65	**13.84**	**7.86**	2.85	**6.64**	**5.58**
Lower (cM)	**32**	**6**	0	**39**	**8.39**	11	**24.54**	**8**
Highest (cM)	**37.82**	**6.05**	13	**44**	**11**	27	**57**	**17**
Upper (cM)	**47**	**6.51**	81	**51**	**14**	72	**65.30**	**29**
LOD	3.99	**6.57**	3.40	1.84	**5.33**	4.97	**5.86**	3.39
Lower (cM)	49	**6**	7.51	0	**13**	52	**21**	3
Highest (cM)	57.95	**6.51**	44.24	49.15	**37**	54	**54.81**	9.80
Upper (cM)	65.60	**11**	93	84.93	**37**	66	**57**	61

Table 3.5 – (previous page) Bayesian 95% confidence intervals of QTLs in the separate analyses of each mapping family. Table includes all chromosomes that were identified as having putative QTLs in the combined analysis of medial forewing band shape. Top shows backcross progeny from B14, bottom shows F2 progeny from B10. Values in bold are from QTLs that were significant in that mapping family.

It appears the QTL on chromosome 15 may potentially be two different QTLs, with one identified in each mapping family. One of these is a broad peak at ~11cM that is well above the significance threshold in both the combined analyses and in the backcross progeny from B14, while a slight rise can also be seen in the F2 progeny from B10 (Figure 3.21). This region corresponds closely markers around a known colour pattern QTL in *Heliconius* that contains the gene *cortex* (HMEL000025) (Nadeau *et al.* 2016). The other is a narrow peak at ~36cM that is only significant in the F2 progeny from B10 ($P < 0.045$), but is also found to be significant and has a higher LOD score than the other QTL in the combined analysis (Figure 3.21).

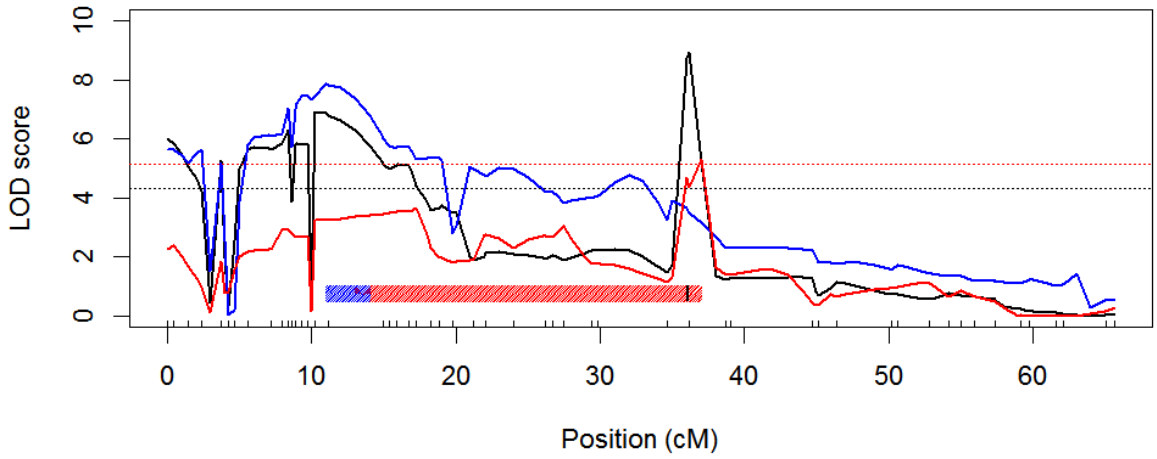


Figure 3.21 – LOD scores across chromosome 15 from the combined family analysis (black), the analysis of F2 progeny from B10 (red) and the analysis of backcross progeny from B14. Dashed lines show genome wide threshold for significant LOD for colour corresponding analysis, not black and blue at approximately same level. Boxes show Bayesian 95% confidence intervals as detailed in Table 3.4 and Table 3.5 for colour corresponding analysis. Markers are identified as tick marks below.

Figure 3.22 – Major effects of the QTL identified on chromosome 10 on medial band shape, at marker identified as having highest LOD score in each mapping family. A) F2 mapping family B10, B) backcross mapping family B14. Elements are numbered from left to right, 1-7.

Modelling the effects of loci found to be significant in each mapping family, revealed the chromosome 10 locus to have the largest effect on medial band shape. These were on the whole fairly consistent between the mapping families, and manifested themselves mainly in changes to the size and shape of elements 1, 5, the distal edge of 6 and the size and shape 7, with smaller effects also visible on the other elements (Figure 3.22). The other loci were generally seen to have smaller but still considerable effects in the backcross mapping family B14. Of particular interest was the locus on chromosome 13 which affected the distal edge of elements 1, 2, 3 and 4, which is consistent with a locus on chromosome 13 called *Ro,* identified in *H. erato* (Nadeau *et al.* 2014) which has also been found to be involved in the rounding of the distal edge of the forewing band. The positions of these two loci also match up, with the *Ro* locus from Nadeau *et al.* (2014) on scaffold Hmel213051 at 49.147cM on my linkage map, the same location as the peak from my QTL analysis. Interestingly, the effects of the first loci on chromosome 15 (at ~11cM) were found to be stronger in the backcross family B14 in which it was significant, than in the F2 family B10, where the second chromosome 15 QTL (at ~37cM) had more effect. The effects of other loci also can be seen in Figure 3.23.

3.4 DISCUSSION

Understanding the genetic basis of diversity is fundamental to evolutionary biology. Theory suggests that one or a few loci should account for most of the variation in any given adaptive walk towards a local phenotypic optimum, with larger effect mutations being substituted earlier and smaller effect mutations evolving subsequently after (Orr 2005b). In *Heliconius* mimicry, a handful of large effect loci have now been found that control major phenotypic differences between subspecies (Merrill *et al.* 2015). These have for the most part been identified in the two species *H. melpomene* and its comimic *H. erato* (Jiggins *et al.* 2005; Baxter *et al.* 2008b; Papa *et al.* 2013; Nadeau *et al.* 2014). Furthermore, it has for the most part been shown that these major effect loci are

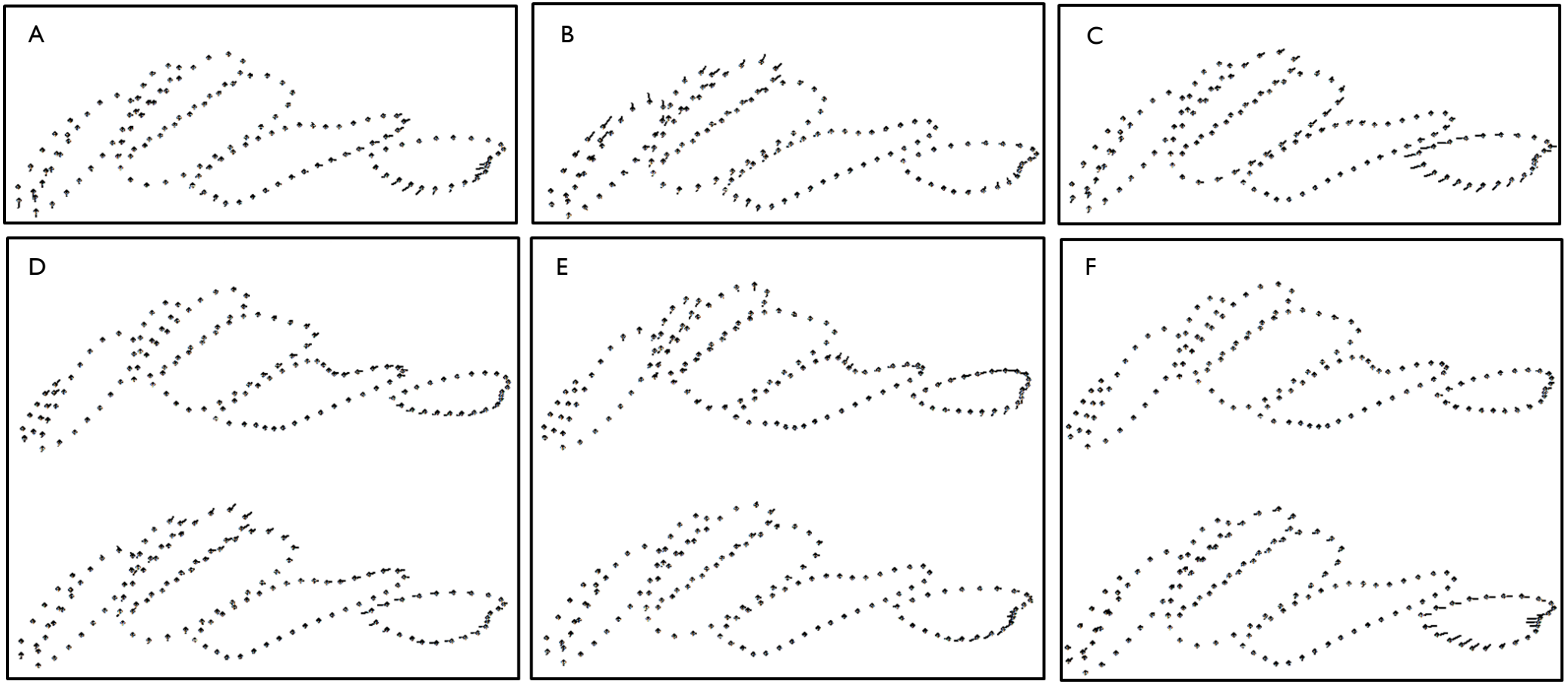


Figure 3.23 – Minor effects of QTLs identified in separate family analyses, on medial band shape, at marker identified as having highest LOD score in each mapping family. A) Chromosome 9 QTL, from backcross family B14; B) Chromosome 13 QTL, from backcross family B14; C) Chromosome 20 QTL, from backcross family B14; D) Chromosome 15 QTL at 11cM, upper - F2 family B10, bottom - backcross family B14; E) Chromosome 15 QTL at 37cM, upper - F2 family B10, bottom - backcross family B14; F) Chromosome 18 QTL, upper - F2 family B10, bottom - backcross family B14.

homologous in between both species, with the same major loci controlling variation in both species (Baxter *et al.* 2008b; Reed *et al.* 2011; Joron *et al.* 2011; Nadeau *et al.* 2016). In this chapter I confirm a major locus named *Ac,* which is known to control the presence of the broken band in *H. erato*, and to affect band shape in *H. melpomene*, plays the same role in the broken band in *H. melpomene*. In addition, I identify the major locus controlling red-orange pigmentation in *H. melpomene*, and explore the minor effects of major effect loci contributing to other phenotypic traits.

3.4.1 *WntA* controls forewing discal melanisation

Through QTL mapping between *H. melpomene aglaope* and *H. melpomene meriana*, I identify that the known colour pattern locus *Ac* explains the majority of phenotypic variation in the binary trait, the broken band. This locus has already been found to control the presence and absence of this trait in *H. erato,* though it is named *Sd* (Martin *et al.* 2012; Papa *et al.* 2013). In addition it had previously been shown to also affect forewing band shape variation in both *H. erato* and in *H. melpomene* (Martin *et al.* 2012). Experiments using Heparin injections and in-situ hybridisations, further supported the theory that the gene *WntA,* a morphogen*,* is most likely gene controlling melanisation in both of these species (Martin *et al.* 2012; Gallant *et al.* 2014b). In the analysis presented here approximate 95% Bayesian confidence intervals placed the QTL at just two markers, Hmel210004_1753431 and Hmel210004_ 1864446, with the highest LOD score at the latter (Hmel210004_ 1864446). The proximity of this marker to the gene *WntA* (HMEL018100), just 5kb upstream, supports the hypothesis that *WntA* also controls the presence or absence of the broken forewing band in *H. melpomene* through melanin patterning.

Nymphalid colour pattern is hypothesised to be determined by an underlying nymphalid groundplan with an array of stereotypical organising centres that control patterning through source-sink relationships causing gradients of different molecules that determine

patterning (Nijhout 1990). This system of organising centres therefore determines the boundaries and positions of colour pattern elements, and creates a modular system in which regulatory changes to genes can affect one part of the pattern while leaving others unaffected (Beldade & Brakefield 2002). In *Heliconius*, evidence for the modularity of cis-regulatory control of patterning genes can be seen in the differential expression of *optix,* that determines different colour elements (Reed *et al.* 2011; Martin *et al.* 2014b; Wallbank *et al.* 2016). *WntA* is a highly conserved morphogen, predicted in *Heliconius* to control colour pattern via outward melanisation from an array of organising centres around the forewing band, each working as a shutter to determine yellow and black pigmented scales (Gilbert 2002; Kronforst *et al.* 2007; Papa *et al.* 2013). *WntA* is located between 1,823,401 – 1,859,103bp on the scaffold Hmel210004, in this linkage map there are no markers within this region, but the closest marker to *WntA*, just 5kb upstream, does have the highest LOD score. However, the next marker upstream (Hmel210004_1903831), just ~40kb away from this peak of LOD, has a LOD score of considerably less, with a drop of LOD 5.35. In contrast the tail in the other side downstream of *WntA* is considerably longer with the next marker being Hmel210004_1753557, ~110kb from the peak. LOD drops off considerably less, with this marker having a LOD score just LOD 0.004 below the maximum. Again, ~46kb further downstream LOD again drops off considerably more. It can therefore be tentatively posited that the functional regulatory region controlling *WntA* expression is likely to be in the ~70kb region downstream of *WntA.*

3.4.2 **Ommochrome pigmentation**

QTL mapping between *H. melpomene aglaope* and *H. melpomene meriana*, identified a locus on chromosome 13 that explains a large proportion of the variance in red-orange pigmentation. This was consistently seen in all analyses, both in the individual mapping families and when these were analysed together. In only one analysis was the clear peak found not to be significant, this was in the analysis using $\log_{10}$ transformed ventral green

values from the backcross progeny only. However, when using the multivariate data for this mapping family, this locus was significant; with the lack of significance in the analysis using $\log_{10}$ transformed ventral green values can be explained by the decreased variation found in the backcross progeny. This result is therefore robust to changes in mapping family identity.

Interestingly a QTL on chromosome 13 involved in forewing band variation, has previously been identified in both *H. melpomene* (Baxter *et al.* 2008a), and in *H. erato* where it has been called *Ro* (Nadeau *et al.* 2014). Based on approximate 95% Bayesian confidence intervals, the location of the QTL identified between *H. melpomene aglaope* and *H. melpomene meriana*, is within a region between 49cM and 60cM, with LOD greatest at the marker Hmel213051_54727. Unfortunately given the broad region this QTL covers, the functional gene at this QTL could be identified. However, the location of this QTL is congruent with the location of the previously identified *Ro* locus (Nadeau *et al.* 2014), which was identified as being on the scaffold Hmel213051, with the nearest gene between 15,332 and 18,649bp, just ~30kb from the LOD peak from my analysis. It seems likely that these two loci, *Ro* and *Or*, are in fact one and the same, with forewing band shape a minor effect of the locus that controls red-orange ommochrome pigmentation in *H. melpomene*. This is further supported from the analysis of forewing medial band shape that also identified this secondary role for *Or* in one of the two mapping families.

3.4.3 **Continued deployment of loci over multiple effects**

This dual effect of loci was one of clear patterns from my QTL analysis. Although some of these results should be treated cautiously, as P-values in some cases were only just significant and do not take multiple testing fully into account. Furthermore, the wide regions that some of these minor effect loci cover, make it possible that what looks like one locus may be multiple loci, each affecting different phenotypes. However despite

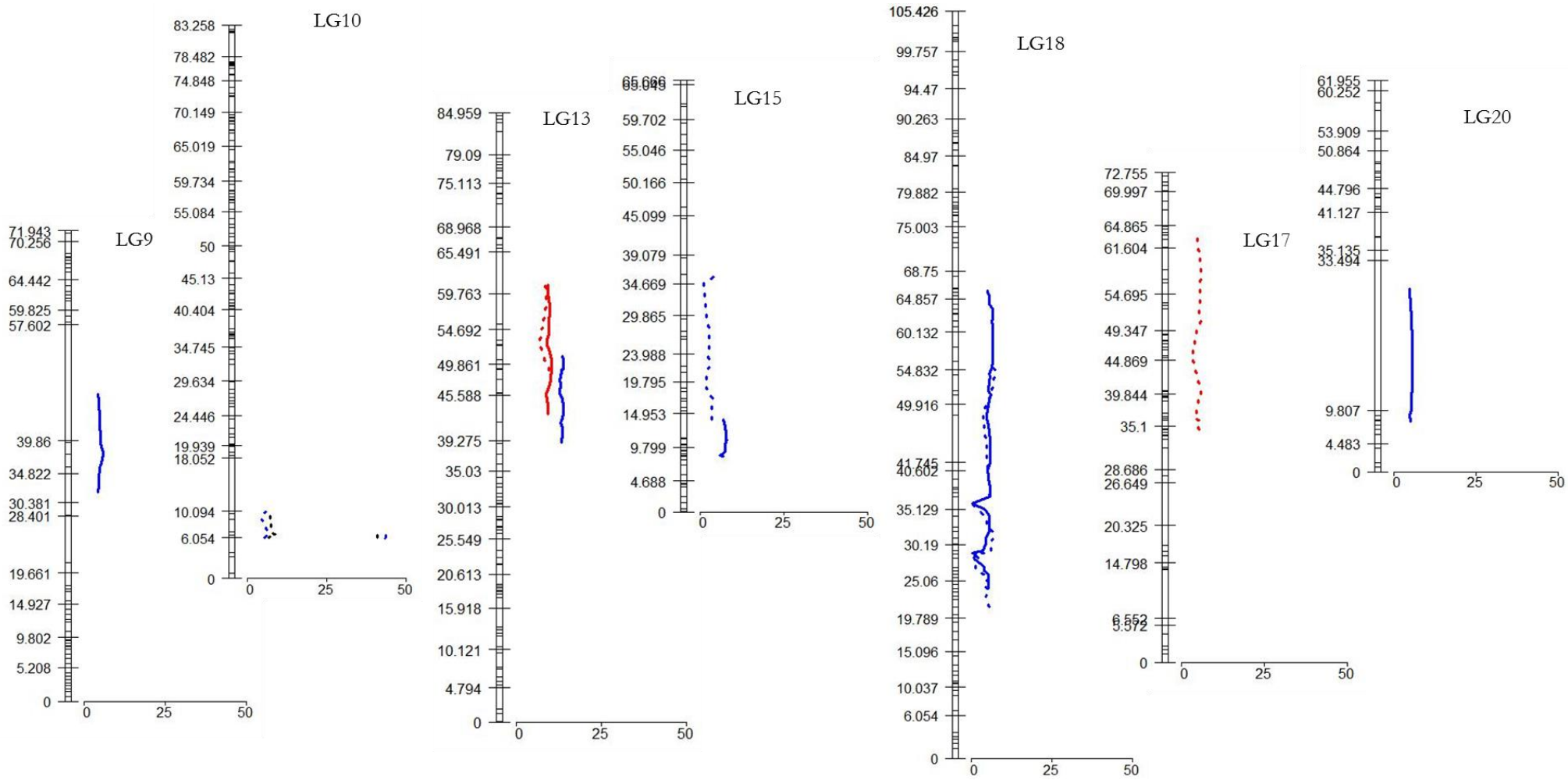


Figure 3.24 – Linkage groups with QTLs (Bayesian intervals) from individual mapping family analyses Positions are given at closest marker to 5cM intervals. X-axes show LOD score. Colours: blue, medial band shape; red, ommochrome pigmentation (multivariate); black, broken band. Dashed lines, B10; solid line, B14.

these caveats, across analyses on different phenotypes I did consistently find loci known or shown to be involved in the control of certain phenotypes, also seemingly involved in the control of others. This was seen both for minor loci and major loci. A good example of the former being the consistent finding of a peak on the latter half (in cM not bps) of chromosome 17. This was predominantly seen in the F2 progeny from B10, and was found to be significant in both the combined family analysis of medial band shape, the multivariate analyses of all $\log_{10}$ transformed RGB scores from B10 progeny, and seen but not significant in this same mapping family in the broken band analysis.

The major locus known to be on chromosomes 15 (*Yb*/*N*) was also found to have minor effects on other phenotypes, for medial band shape this was seen in both mapping families, with a significant QTL found at *Yb*/*N*, as well as a possible second found upstream from this (Figure 3.21). A QTL was also found in this rough region in the multivariate analysis on red-orange pigmentation from the combined mapping families, while in the multivariate analysis of red-orange pigmentation on the backcross progeny from B14 an almost significant peak was found in the *BD* region. For both mapping families significant peaks on chromosome 18 were also found to explain variation in medial band shape, although this broad region did not necessarily appear to cover the *BD* locus. This pattern, although patchy due to differences in mapping families, is consistent with previous research in *Heliconius*, which has found that the regions on these chromosomes play a role in determining patterns of melanisation across elements, especially those in which they are involved in patterning (Baxter *et al.* 2008a; Huber *et al.* 2015). This is perhaps not surprising given the strategic placement of these patterning genes (*cortex* and *optix*) in developmental pathways involved in wing patterning, which control downstream processes involved in pigmentation and scale structure (Merrill *et al.* 2015). Given the potential dual effects conferred by the *Ro*/*Or* locus in *H. melpomene*, it can be further supposed that if this is a single gene, it may also have a similar effect and placement in butterfly wing patterning pathways.

4. Modulation and introgression of mimicry elements

4.1 INTRODUCTION

The convergent evolution of similar phenotypes between species can be driven by a number of different processes (Stern 2013). The most common of these is independent evolution, in which similar phenotypes evolve independently in each species. This can either be through different changes to the same or even different genetic and developmental pathways (Gross *et al.* 2009), or through independent convergent genetic evolution where mutational changes effect the same genes in the same genetic and developmental pathways (Zhang 2006; Tishkoff *et al.* 2007; Gompel & Prud'homme 2009; Parker *et al.* 2013). Alternatively, convergent phenotypic evolution can be driven by 'collateral evolution' (Stern 2013), another form of convergent genetic evolution. This occurs when alleles are shared among populations or species, either through shared ancestry and incomplete lineage sorting, or through the introgression of these alleles from one population/species to another (Stern 2013).

Introgression occurs through hybridization and subsequent backcrossing between individuals from different species (Twyford & Ennos 2012). This has the power to be a creative evolutionary process, allowing advantageous alleles and adaptive allelic combinations to accumulate faster than by mutation alone. This can potentially drive speciation, if the traits that introgress are involved in mate choice. However, evidence of adaptive introgression is rare with only a few examples known, such as insecticide resistance in *Anopholes* mosquitoes; warfarin resistance in mice (Song *et al.* 2011);

between ancient humans, Neanderthals and Denisovans (Racimo *et al.* 2015); in the tunicate *Ciona intestinalis* (Roux *et al.* 2013) and in Darwin's finches (Lamichhaney *et al.* 2015) as well as handful of other potential examples (Hedrick 2013). These examples are few and far between, and in some cases the evidence is somewhat circumstantial. In addition, in many of these cases, such as those of mice, mosquitoes, and coyote and wolf coat colour (Hedrick 2013), human disturbance and influence is implicated in causing or assisting this process.

Heliconius butterflies have bright, aposematic and mimetic colour patterns that are hotspots for the repeated convergent evolution of adaptive phenotypic variation across species (Baxter *et al.* 2008b). In *Heliconius,* mimicry also leads to phenotypic divergence within species, with colour pattern loci easily identifiable in population genomic studies as clear islands of divergence across intraspecific hybrid zones (Baxter *et al.* 2010; Counterman *et al.* 2010; Nadeau *et al.* 2013, 2014). Phylogenetic discordance between these colour pattern loci and the rest of the genome, as well as elevated levels of shared derived sites at these, indicates that hybridization has led to the adaptive introgression of colour pattern loci between several different species; *H. melpomene, H. elevatus* and *H. timareta*, all of which are within the wider *H. melpomene*-silvaniform clade (Dasmahapatra *et al.* 2012; Nadeau *et al.* 2013; Pardo-Diaz *et al.* 2014). This means that in *H. melpomene* and some of their silvaniform comimics, phylogenies at colour pattern loci often group taxa by colour pattern phenotype rather than by species or geographic proximity (Dasmahapatra *et al.* 2012; Pardo-Diaz *et al.* 2012; Wallbank *et al.* 2016). This is due to strong selection on colour pattern, coupled with a history of shared ancestry of these loci across these species, either through adaptive introgression or shared ancestral polymorphism (Martin *et al.* 2014a).

In this chapter I take advantage of this shared ancestry of colour patterns between *H. elevatus* and *H. melpomene*, and the clear and narrow signal of divergence across

intraspecific hybrid zones (Figure 4.1), to identify loci in these two species that are both shared and derived, relative to *H. melpomene* with other colour pattern phenotypes, and to *H. pardalinus butleri*, the sister species of *H. elevatus*. It is important to note here, that previous work has shown a history of gene flow between *H. pardalinus* and *H. elevatus* despite their vastly different colour patterns, with *Fst* generally only high at known colour pattern loci (Kryvokhyzha 2014). This conserved signal of shared ancestry between taxa of different species, but that match colour pattern phenotypes, indicates a region under mimicry selection, and involved in colour pattern control (Martin *et al.* 2014a). On the other hand, across most other genomic regions not under mimicry selection, gene flow which is greater among species, than between, will tend to homogenise the genome (Wu 2001). This homogenisation across the genome except for at colour pattern loci will be clearest among taxa from/or near intraspecific hybrid zones, leaving a clear signal of colour pattern loci as islands of divergence (Baxter *et al.* 2010; Counterman *et al.* 2010; Nadeau *et al.* 2012, 2014).

The *BD* locus controls red patterning in an array of *Heliconius* species (Jiggins *et al.* 2005; Baxter *et al.* 2008b; Papa *et al.* 2013). This is through regulation of a gene *optix*. In *H. melpomene*, *H. erato, H. cydno*, as well as in *H. doris* and *H. atthis* expression of this gene during development has been shown to prefigure red patterning (Reed *et al.* 2011; Martin *et al.* 2014b). The signal of shared ancestry across species, combined with recombination breakpoint analysis has already been used to define two regulatory modules in the *H. melpomene* clade thought to control expression of *optix* and the presence and absence of the rays and dennis pattern elements (Wallbank *et al.* 2016). Furthermore, this study also looked at the complex evolutionary phylogenies of these regulatory subunits, and suggested different origins for each, with the rays introgressing from *H. melpomene* into *H. elevatus*, while the dennis patch introgressed from *H. elevatus* into *H. melpomene* (Wallbank *et al.* 2016).

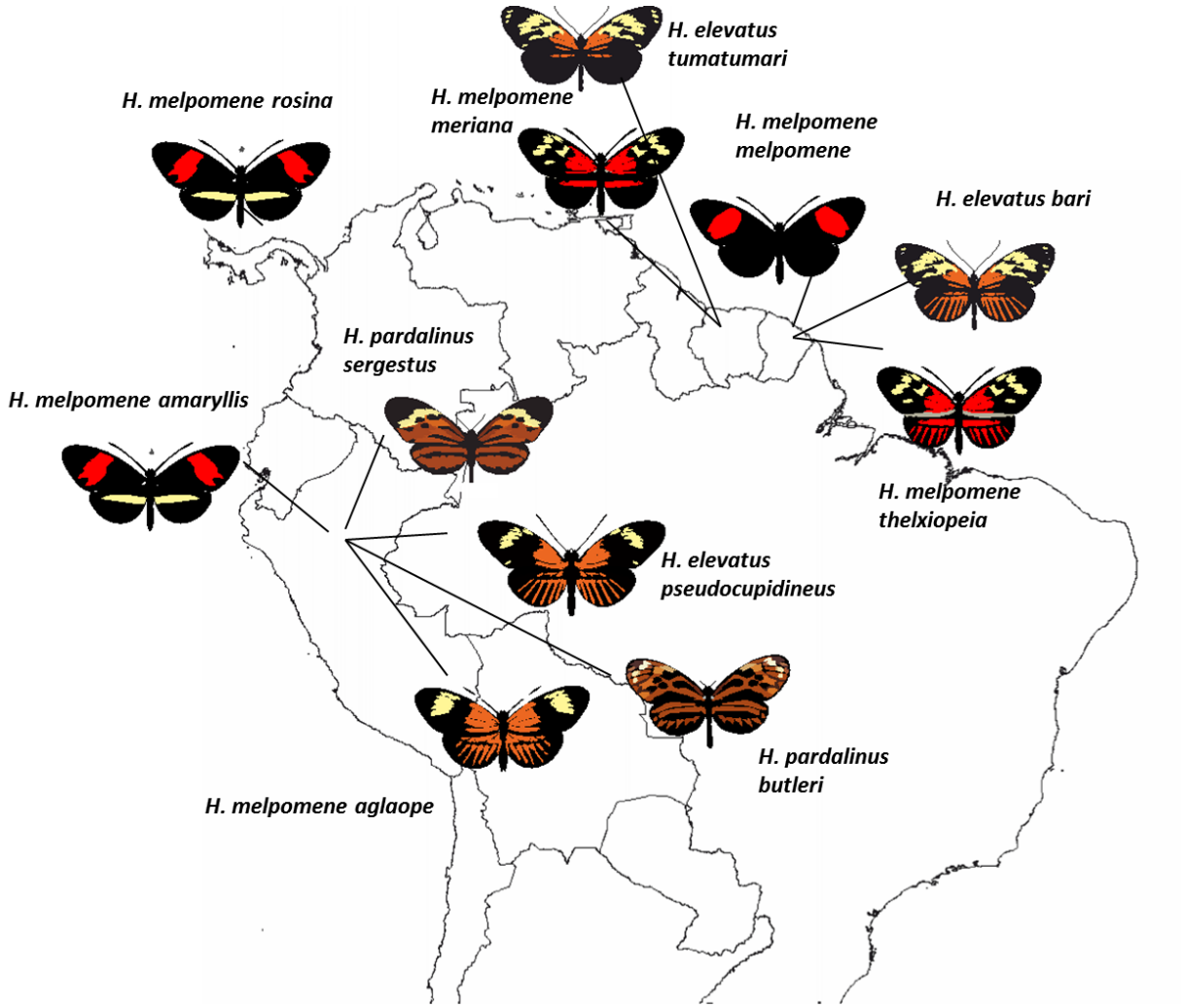


Figure 4.1 – Colour pattern races of *H. pardalinus*, *H. elevatus* and *H. melpomene* used to identify loci shared and derived between species with matching colour patterns from two *H. melpomene* hybrid zones.

Two other genes have also been found to be important in wing patterning, in *Heliconius* butterflies. These are the *WntA* gene involved in melanic patterning across the discal portion of the forewing (Martin *et al.* 2012; Gallant *et al.* 2014a; Kronforst & Papa 2015), and *cortex* that is involved in the control of yellow wing pattern elements (Nadeau *et al.* 2016). In this chapter, as well as defining regulatory modules, and looking for shared defined regions around these wing patterning genes I investigate phylogenetic discordance across these regions at a finer scale than has previously been done, and find evidence of an even more complex exchange of mimicry loci, between *H. melpomene* and *H. elevatus*. Furthermore, while studies have so far utilised multiple races of *H. melpomene* from different geographic regions with different colour patterns, they have only included rayed

forms of *H. elevatus* from Peru. In Guyana and western Suriname, *H. elevatus tumatumari* lacks the rays element found in colour pattern races from the rest of the Amazon basin (Brown *et al.* 1974). In addition, both *H. elevatus tumatumari* and *H. elevatus bari* from French Guiana, and their *H. melpomene* co-mimics all have split forewing bands. I use multiple races of *H. elevatus* and *H. melpomene* from the hybrid zones in this region, as both a second independent test for delimiting wing colour pattern loci, and to identify whether the introgression between comimics in the Guianas has led matching patterns or whether this has been driven by independent genetic evolution.

4.2 METHODS

4.2.1 **Sample collection and seqencing**

Two *H. elevatus tumatumari* samples, and two *H. pardalinus butleri* samples, and one *H. elevatus bari* sample (see appendix 1 for details) were sequenced to >30x idealised coverage. RNA-free genomic DNA was extracted to a concentration of approximately 15ng/µl from thoracic tissue using a Qiagen DNeasy Blood and Tissue Kit following the standard protocol provided by the manufacturer. Libraries were prepared using TruSeq DNA PCR-Free Library Preparation Kits (by K Dasmahaptra), with an insert size of approximately 350bp. Libraries were sequenced to ~30x coverage on an Illumina HiSeq 2000 instrument at the FAS Center for Systems Biology.

In addition to these whole-genome shotgun sequenced paired-end libraries were available (from ENA; http://www.ebi.ac.uk/ena/). These were from four *H. melpomene aglaope* (ERS235655, ERS235656, ERS235657, ERS235658), two *H. melpomene amaryllis* (ERS235651, ERS235654), one *H. melpomene thelxiopeia* (ERS977708), two *H. melpomene meriana* (ERS977704, ERS977703), four *H. melpomene rosina* (ERS074426, ERS235641, ERS235642, ERS235643), four *H. elevatus pseudocupidineus* (ERS070236, ERS977673, ERS977674, ERS070238), three *H. elevatus bari* (ERS977670, ERS977671, ERS977672,

xxxxxx), four *H. pardalinus sergestus* (ERS074426, ERS235641, ERS235642, ERS235643,), one *H. ethilla aerotome* (ERS977677) and two *H. hecale felix* (ERS977681, ERS235670) (see appendix 1 for details).

4.2.2 Variant calling

Reads from silvaniform taxa were mapped to the *H. melpomene* reference genome v2 (Davey *et al.* 2016) using Stampy1.0.27 with a substitution rate of 0.05. *H .melpomene* samples were aligned with BWA. BAM files were then sorted, duplicate reads were removed, GATKs HaplotypeCaller was then used to call SNPs with the parameters -out_mode EMIT_ALL_CONFIDENT_SITES, -baq CALCULATE_AS_NECESSARY, -hets 0.01 and -emitRefConfidence GVCF. GVCFs were then combined and genotyped with –CombineGVCFs and GenotypeGVCFs (from GATK). Bcftools v1.3.1, was then used to filter GVCFs with minimum read depth set to 5 and Genotype Quality set to 30. Beagle 4 was then used to infer phasing and impute missing data with the following parameters; impute=true, nthreads=15, window=10000, overlap=1000 and gprobs=false. Python scripts from Simon Martin (available at https://github.com/simonhmartin) were then used to Parse this VCF to make a phased genotype calls file.

4.2.3 Phylogenetic weighting method

In order to identify shared putative regulatory regions that control expression of major colour pattern genes (*cortex*, *optix* and *WntA*), I employed a descriptive, phylogenetic weighting method, called Topology Weighting by Iterative Sampling of Subtrees; Twisst (available from: https://github.com/simonhmartin/twisst). This method can quantify the phylogenetic relationships among taxa in narrow regions across the genome. This is done by sampling trees, with one sample representing each specified taxon. The proportion of trees of each topology type is then calculated as that topology weighting. This is done in sliding windows across a genomic region. In doing so, it can identify regions of the genome with different phylogenetic histories to the null expectation; the species tree.

Using comparisons with taxa from two natural colour pattern hybrid zones of *H. melpomene* (postman-rayed in Peru; postman-rayed-nonrayed in the Guianas), along with sympatric races of the silvaniform *H. elevatus* (a co-mimic of rayed *H. melpomene*) and its sister species *H. pardalinus*, I use this phylogenetic weighting method to identify putative colour pattern control modules that show shared ancestry across species that share colour pattern phenotypes, a method analogous to 'phylogenetic footprinting' (Cliften 2003).

4.2.4 **Pairwise phylogenetic comparisons**

PhyML 3.0 was used to produce neighbour joining trees that included all samples, from 3kb genomic windows across the three chromosomes of interest. This size window reduces noise but is still narrow enough not to swamp signal. Weightings were estimated using a dynamic threshold, such that trees are sampled until the 95% binomial confidence interval around each weighting was less than 5%. An experimental design, that I term here a 'pairwise phylogenetic comparison', was used to identify putative regulatory regions controlling colour pattern. This was done through two Twisst comparisons, with five taxa in each, and a total of six taxa across both. Three of these were orienting taxa (included across both runs) and three were focal taxa.

In the first Twisst comparison, the two focal taxa were the primary (Figure 4.2 taxa A) and secondary taxa (Figure 4.2 taxa E). These are taxa of two different species that in general share the colour pattern phenotype of interest. Trees in most windows are expected to show one of the three possible species trees. These species trees group silvaniform taxa and *H. melpomene* taxa separately (Figure 4.2 top row). However, if trees group the primary and secondary taxa together, with one orienting taxon basal to this group, this suggests shared ancestry between the primary and secondary taxa (Figure 4.2: 1 and 2, top middle and bottom rows). Thus, weightings (the proportion) of trees showing shared ancestry between the primary and secondary taxa were calculated in each

3kb window. In the second Twisst comparison, weightings of trees consistent with a hypothesis of shared ancestry were calculated between the primary taxon and a tertiary taxon (Figure 4.2 taxa F). Again this tertiary taxa was of a different species to the primary taxa, but closely related to the secondary taxa.

Through these two comparisons, regions with shared ancestry between the primary and secondary taxa (generally with shared colour pattern phenotypes), but divergent between the primary and tertiary taxa could then be identified. This was done by subtracting the shared ancestry signal from the primary and tertiary taxa (comparison two) from the shared ancestry signal between the primary and secondary taxa (comparison one). A pairwise weighting score of 0 then suggests equal shared ancestry between the primary and secondary, and the primary and tertiary taxa. This can either be through none or all of topologies showing shared ancestry between both the primary and secondary taxa, and the primary and tertiary taxa. In contrast, a negative pairwise weighting indicates shared derived ancestry between the primary and tertiary taxa, while a positive pairwise weighting indicates shared derived ancestry between the primary and secondary taxa.

It is important to note that the primary taxon must be a different species to that of the secondary and tertiary taxa, otherwise simply the null expectation of the species tree will be seen. Pairwise phylogenetic comparisons in which the primary, secondary and tertiary taxa all share the same colour pattern phenotype were also carried out. In these the expectation is that there should be no difference in the signal of shared ancestry, between the primary and secondary, or primary and tertiary taxa. This serves as a control for comparisons in which colour pattern phenotypes do differ.

The three orienting taxa used were kept constant in all comparisons. These were a Panamanian *H. melpomene* race *H. m rosina* (allopatric from the primary, secondary or tertiary taxa), and two other silvaniform species (*H. hecale* and *H. ethilla*). This choice of orienting taxa results in an expected 'root' placed somewhere along the branch separating

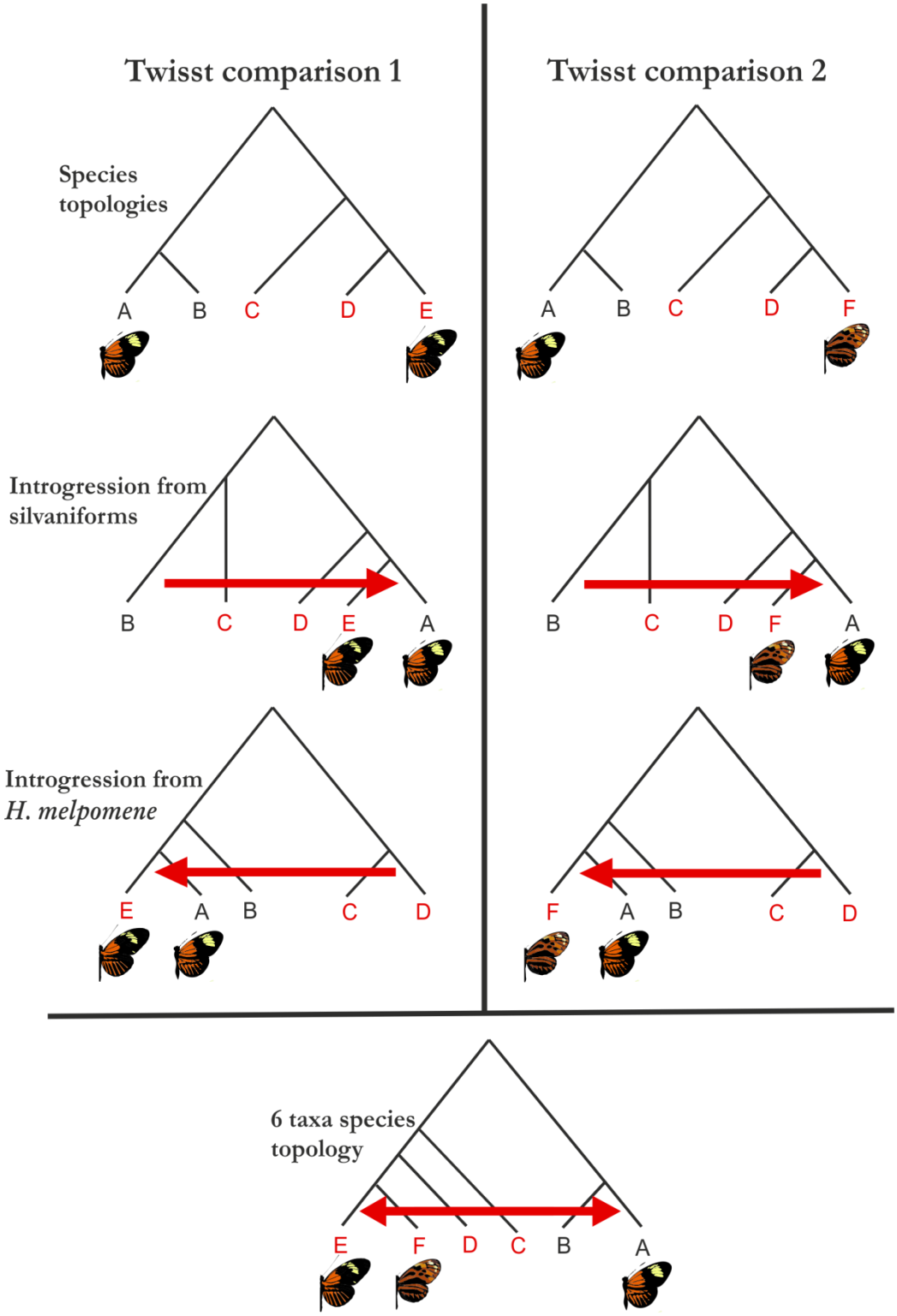


Figure 4.2 – Pairwise phylogenetic comparison design, with two Twisst comparisons. Shows example with *H. m. aglaope*, *H. e. pseudocupidineus* and *H. p. butleri*. Each letter represents a taxa, red are silvaniform taxa, black are *H. melpomene* taxa. Red arrows show possible introgression. Bottom row shows species tree with all six taxa, equivalent to pairwise phylogenetic comparison, and showing shared derived introgression calculated from pairwise design. A) primary taxa, E) secondary taxa, F) tertiary taxa. Other taxa (B, C, D) are orienting taxa, *H. m. rosina*, *H. ethila* and *H. hecale*.

the silvaniform taxa from the *H. melpomene* taxa. Including these taxa also allow the direction of introgression to be ascertained. If the taxon basal to the clade of focal taxa is silvaniform, this suggest introgression from *H. melpomene* (Figure 4.2: 1 and 2, bottom row), while if the taxa basal to the clade of focal taxa is *H. melpomene rosina*, this suggest introgression from a silvaniform (Figure 4.2: 1 and 2, middle row). The inclusion of two silvaniforms also helps control for introgression that may have occurred between other non-focal silvaniform taxa and *H. melpomene* taxa. The choice of the allopatric *H. melpomene rosina* race, means that any shared derived regions between other *H. melpomene* taxa and *H. elevatus*/*H. pardalinus* taxa, have occurred since the focal *H. melpomene* taxa split from *H. melpomene rosina*. The most likely explanation for this is therefore one of adaptive introgression of colour pattern loci between sympatric races of *H. melpomene* and *H. elevatus* that share colour pattern phenotypes.

This pairwise design is equivalent to running a single phylogenetic comparison that includes all six taxa (Figure 4.2: 3). However, including only five taxa in each separate phylogenetic comparison keeps the total possible number of unrooted topologies to 15, rather than the 105 given by just six taxa. This keeps topologies simple while at the same time allows the ancestral source of shared loci (direction of introgression) to be identified. While single phylogenetic comparisons using just five taxa, for example with *H. pardalinus*, a *H. melpomene* and a *H. elevatus* comimic included, can potentially get at the same answer, the pairwise design also reduces noise by including an additional non-gene flow species.

4.2.5 **Non-pairwise phylogenetic comparisons**

In addition to this pairwise design, some phylogenetic comparisons with Twisst were also carried out and analysed independently of any others. These were carried out for two reasons. The first was to identify the ancestral source of shared loci. These comparisons included the three orienting taxa, along with a focal *H. melpomene* and *H. elevatus* race

that shared a colour pattern phenotype. If *H. elevatus* is found within a clade that includes the two *H. melpomene* taxa, this tells us the ancestral source is from *H. melpomene*. Likewise if the inverse if found, with the *H. melpomene* found in a clade that includes the focal *H. melpomene* and *H. elevatus* races and a silvaniform taxa, with *H. melpomene rosina* basal to this, then this indicates a silvaniform ancestral source for this shared locus.

The second reason was to test for shared ancestry between Guianese *H. melpomene* and *H. elevatus* comimics, relative to the Peruvian *H. melpomene aglaope*. These included three taxa, along with the two silvaniform taxa. For some phenotypes that were identical in both dennis only and dennis-rayed Guianese forms of *H. melpomene* and *H .elevatus*, these forms were treated as single taxon. If shared derived ancestry was found between Guianese *H. melpomene* and *H. elevatus* comimics, this would suggest that gene flow occurred between these taxa at this locus since the split between the Guianese forms of *H. melpomene* and *H. melpomene aglaope* from Peru. In contrast if there appears to be shared ancestry between Guianese forms of *H. elevatus* and Peruvian *H. melpomene aglaope*, this suggests that these *H. elevatus* loci are derived from the Peruvian *H. elevatus* loci.

4.2.6 **Plotting shared ancestry**

Plots of shared ancestry were made in R v3.3.1. Plots across whole chromosomes were made using a Loess regression smoothing method with a span of 0.006. This smoothing allows easier visualisation by increases the signal to noise ratio across larger regions (code for implementing this Loess smoothing algorithm is available from https://github.com/simonhmartin/twisst). Plots showing shared derived loci near colour pattern loci were also plotted in R, without Loess smoothing so that the true signal could be seen for regulatory module identification. In plots of non-pairwise phylogenetic comparisons, two lines are plotted. Zero suggests no shared ancestry, while weighting

scores departing from this indicate shared ancestry between relevant taxa, or in directional plots the ancestral source of the shared derived allele.

4.3 RESULTS

4.3.1 **Regulatory control of *optix* expression**

Pairwise phylogenetic comparisons across both Guianese and Peruvian taxa, consistently identified a number of regions showing a shared and derived ancestry between *H. melpomene* and *H. elevatus* races with matching colour pattern phenotypes upstream of the transcription factor *optix*. These regions are putatively described as a *rays*, *dennis* and *band* locus. These were investigated with eight pairwise phylogenetic comparisons in which the focal taxa were (primary, secondary and tertiary): i) *H. melpomene aglaope*, *H. elevatus pseudocupidineus*, *H. pardalinus butleri*; ii) *H. elevatus pseudocupidineus, H. melpomene aglaope*, *H. melpomene amaryllis*; iii) *H. pardalinus butleri, H. melpomene aglaope*, *H. melpomene amaryllis*; iv) *H. pardalinus sergestus, H. melpomene aglaope*, *H. melpomene amaryllis*; v) *H. elevatus tumatumari, H. melpomene thelxiopeia, H. melpomene melpomene*; vi) *H. elevatus bari, H. melpomene thelxiopeia, H. melpomene melpomene;* vii) *H. elevatus bari, H. melpomene thelxiopeia, H. melpomene meriana;* viii) *H. elevatus tumatumari, H. melpomene meriana, H. melpomene thelxiopeia*.

4.3.1.1 The rays *locus*

Pairwise phylogenetic comparisons revealed what appear to be two loci within a ~25kb region associated with the hindwing rays phenotype. Previously this *rays* locus had been defined using recombination breakpoints as a ~37kb region (Wallbank *et al.* 2016). Across this whole *rays* region nearly all trees indicate shared ancestry between the rayed co-mimics *H. elevatus pseudocupidineus* and *H. melpomene aglaope* (Figure 4.3; Figure 4.4a), but not between *H. melpomene aglaope* and *H. pardalinus butleri* which lacks hindwing rays but has dennis (Figure 4.4a; Figure 4.4c). Pairwise phylogenetic comparisons with Guianese

taxa again show this region has a shared ancestry between rayed *H. elevatus bari* and *H. melpomene thelxiopeia*, relative to non-rayed *H. melpomene meriana* or *H. melpomene melpomene* (Figure 4.4f; Figure 4.4g). This region of shared derived ancestry was therefore highly conserved across rayed taxa. Furthermore, across this whole region comparisons with Guianese non-rayed *H. elevatus tumatumari*, do not share ancestry with the non-rayed *H. melpomene meriana,* but instead, this region shows a strong signal of shared ancestry with the rayed forms of *H. melpomene* (Figure 4.4e; Figure 4.4h). This suggests an independent origin for the loss of rays in *H. elevatus tumatumari* from a rayed ancestral form, while *H. melpomene meriana* has likely lost the rays through recombination with the its parapatric conspecific the non-rayed postman *H. melpomene melpomene*.

Surprisingly one of these two, rays loci (Figure 4.4, *rays2*), was also shared between *H. elevatus aglaope* and the non-rayed *H. pardalinus sergestus*, relative to *H. melpomene amaryllis* (Figure 4.4d). This unexpected pattern was further supported, when the ancestral sources of these two rays loci were examined, with this shared *rays2* locus an ancestral Silvaniform allele, while the *rays1* locus appears to ancestrally be from *H. melpomene* (Figure 4.5a). This was what was previously found from a phylogeny built from the *rays2* locus, when it was previously seen as just part of the whole ~37kb *rays* region (Wallbank *et al.* 2016). The exact roles, each of these loci play in controlling the hindwing rays phenotype, are unclear. One possibility is that this independent history suggests that only the ancestral *H. melpomene rays1* locus is functionally important, with the ancestrally silvaniform *rays2* allele simply hitchhiking with the *H. melpomene rays1* region. If this were the case, this narrows the *rays* region down to just 9kb. Alternatively, the modern rays phenotype in these taxa has been constructed from two tightly linked modules with previously independent evolutionary histories.

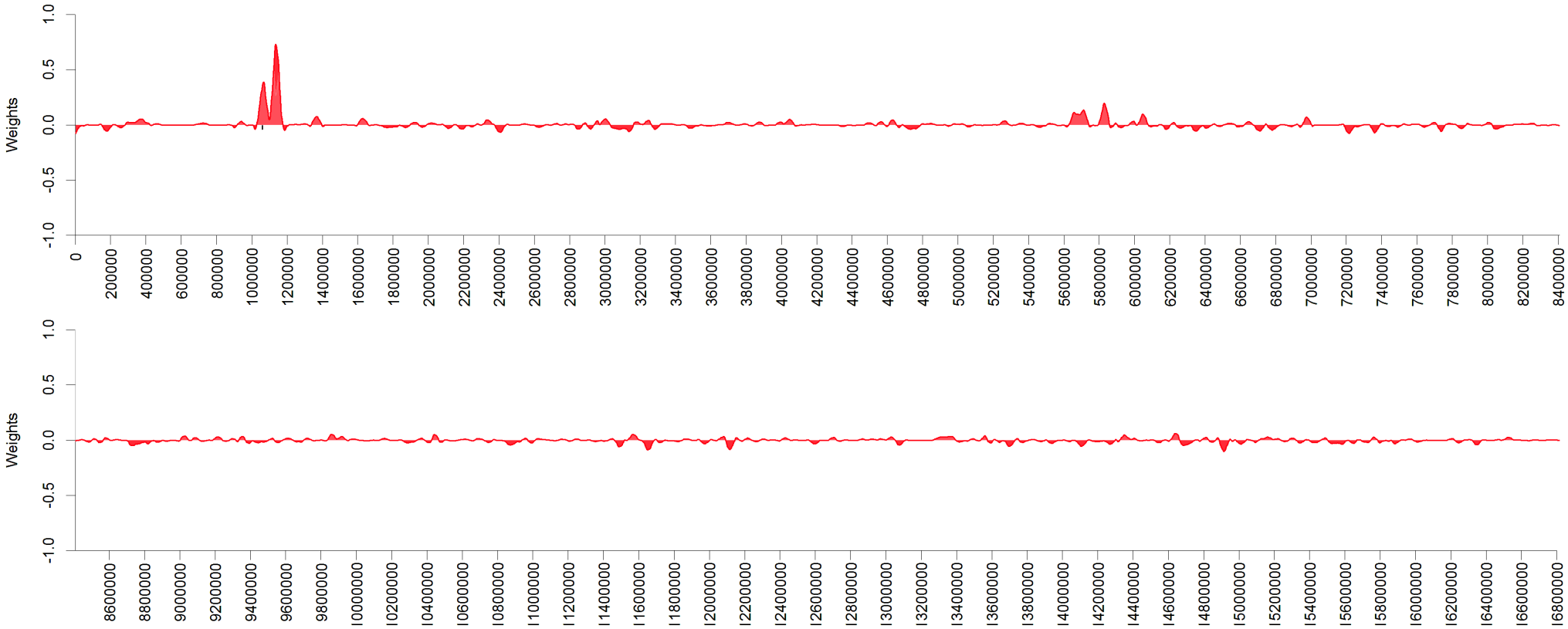


Figure 4.3 – Shared ancestry as Loess-smoothed topology weighting, across chromosome 18 in the pairwise comparison using *H. melpomene aglaope*, *H. elevatus pseudocupidineus* and *H. pardalinus butleri*. Positive values shows shared ancestry between comimics *H. melpomene aglaope* and *H. elevatus pseudocupidineus*, negative values shows shared ancestry between *H. melpomene aglaope* and *H. pardalinus butleri*. Y-axis shows position in base pairs across the chromosome.

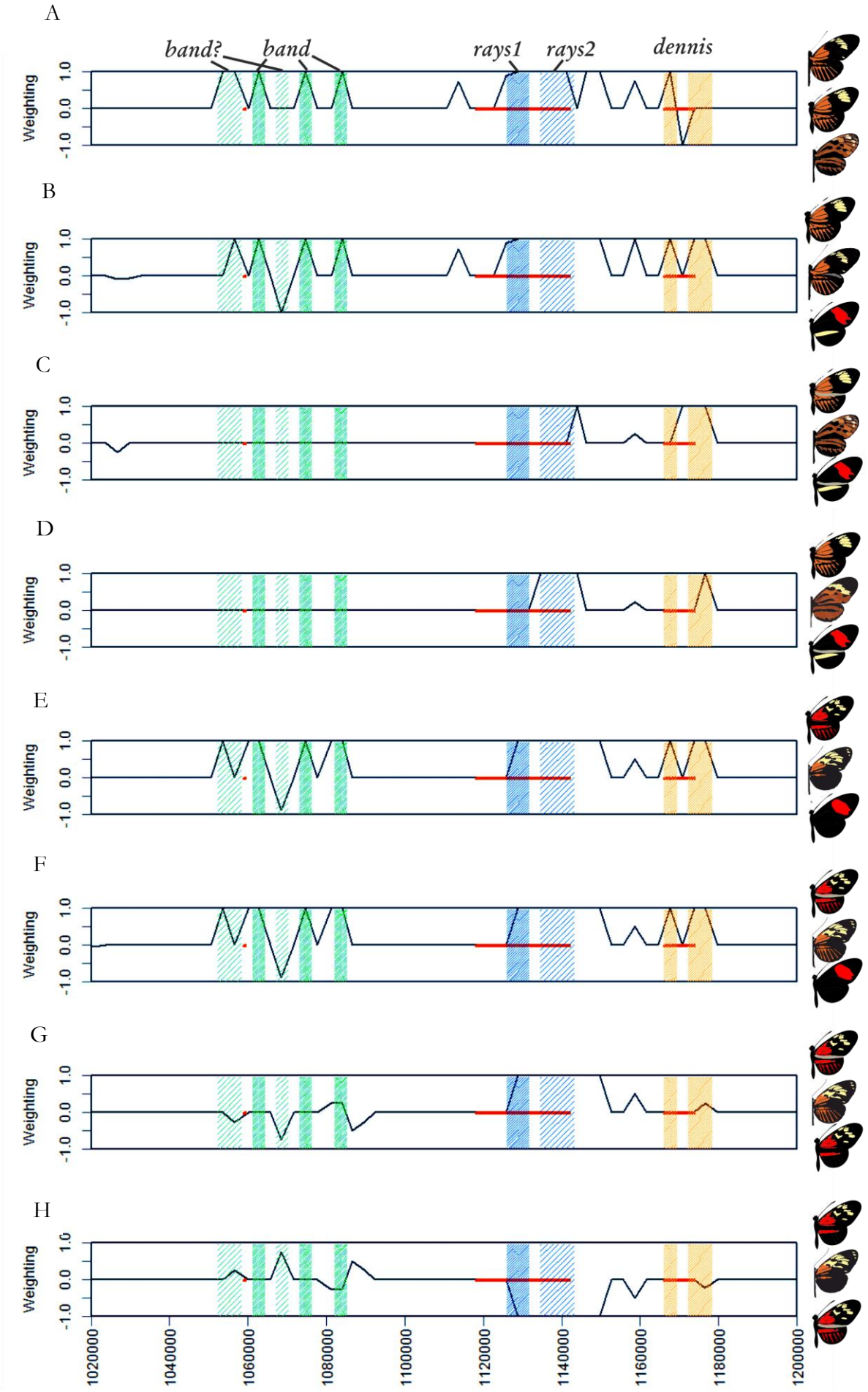

A
band?
band
rays1
rays2
dennis
Weighting
1.0
0.0
-1.0
B
C
D
E
F
G
H
1020000
1040000
1060000
1080000
1100000
1120000
1140000
1160000
1180000
1200000

Figure 4.4 – Pairwise phylogenetic comparisons with shared derived ancestry as topology weightings (proportions) around the *optix* gene. Primary taxa is in the middle and given here first, then top taxa, then bottom taxa. A) Comparison between *H. melpomene aglaope*, *H. elevatus pseudocupidineus*, *H. pardalinus butleri*; B) *H. elevatus pseudocupidineus, H. melpomene aglaope*, *H. melpomene amaryllis*; C) *H. pardalinus butleri, H. melpomene aglaope*, *H. melpomene amaryllis*; D) *H. elevatus tumatumari, H. melpomene melpomene*, *H. melpomene thelxiopeia*; E) *H. elevatus bari, H. melpomene thelxiopeia, H. melpomene melpomene;* F) *H. elevatus bari, H. melpomene thelxiopeia, H. melpomene meriana;* G) *H. elevatus tumatumari, H. melpomene meriana, H. melpomene thelxiopeia*; H) *H. pardalinus sergestus, H. melpomene aglaope*, *H. melpomene amaryllis*. Positive values show shared ancestry between middle (primary) taxa and top taxa, relative to bottom, negative values show shared ancestry between middle (primary) and bottom taxa, relative to top. Red shows from left to right, *optix*, and the previously defined *rays* and *dennis* regions. Shading: Green shows putative *band* modules, blue shows putative *rays* modules, orange shows putative *dennis* modules.

Topology weightings from phylogenetic comparisons that included both Guianese dennis-rayed *H. elevatus bari and H. melpomene thelxiopeia* and/or dennis only *H. elevatus tumatumari* and *H. melpomene meriana* taxa, along with *H. melpomene aglaope* from Peru were also investigated (Figure 4.5). This confirmed the difference in the wider *rays* region, between the non-rayed forms *H. elevatus tumatumari* and *H. melpomene meriana*, with the *rays* locus in non-rayed *H. elevatus tumatumari* more closely related to allopatric rayed *H. melpomene aglaope* than its non-rayed co-mimic *H. melpomene meriana* (Figure 4.5b). Perhaps more surprisingly given this, was the finding of shared ancestry across this region, between the two Guianese rayed forms; *H. elevatus bari* and *H. melpomene thelxiopeia* (Figure 4.5c). This suggests gene flow does or has occurred between these two taxa, since they have diverged from the *H. melpomene aglaope*/*H elevatus pseudocupidineus* alleles.

	Primary	Secondary	Tertiary	*band*	*rays1*	*rays2*	*dennis*
A	*H. m. aglaope*	*H. e. pseudo.*	*H. p. butleri*	Yes	Yes	Yes	Yes
B	*H. e. pseudo.*	*H. m. aglaope*	*H. m. amryllis*	Yes	Yes	Yes	Yes
C	*H. p. butleri*	*H. m. aglaope*	*H. m. amryllis*	No	No	No	Part
D	*H. p. sergetus*	*H. m. aglaope*	*H. m. amryllis*	No	No	Yes	Part
E	*H. e. tumatumari*	*H. m. thelxiopeia*	*H. m. melpomene*	Yes	Yes	Yes	Yes
F	*H. e. bari*	*H. m. thelxiopeia*	*H. m. melpomene*	Yes	Yes	Yes	Yes
G	*H. e. bari*	*H. m. thelxiopeia*	*H. m. meriana*	No	Yes	Yes	No
H	*H. e. tumatumari*	*H. m. meriana*	*H. m. thelxiopeia*	No	Yes	Yes	No

Table 4.1 – Table summarising putative regulatory modules of *optix*, found to be shared derived between primary and secondary taxa, relative to tertiary taxa, from pairwise phylogenetic comparisons from Figure 4.4. For *dennis* where more than one putative loci was identified, part means that one locus shows shared derived ancestry while the other does not.

4.3.1.2 The dennis *locus*

In Wallbank *et al* (2016) a ~7kb region from (813,000 – 820,000bp on Hmel218003) was identified through recombination breakpoint analysis, as being associated with the hindwing and forewing dennis. Comparisons using Peruvian taxa from *H. melpomene* and *H. elevatus* shows two putative regions with shared ancestry between *H. elevatus pseudocupidineus* and *H. melpomene aglaope*, relative to *H. melpomene amaryllis* which does not have the dennis phenotype (Figure 4.4b). In pairwise phylogenetic comparisons with Guianese taxa these regions also show shared ancestry between *H. elevatus bari*, *H. elevatus H. elevatus tumatumari*, *H. melpomene thelxiopeia* and *H. melpomene meriana*, that all share these dennis phenotypes, relative to *H. melpomene melpomene* that does not (Figure 4.4e; Figure 4.4f). This supports what was found in comparisons between Peruvian taxa. These windows are 3kb and 6 kb in size and separated by a 3kb window that instead groups taxa by species. Both of these windows overlap with the 7kb dennis region previously defined by breakpoint analysis, and are shared across all races of Guianese *H. elevatus* and *H. melpomene* that have the dennis (Figure 4.4g; Figure 4.4h).

Interestingly, pairwise phylogenetic comparisons with *H. pardalinus butleri* and *H. pardalinus sergestus*, whose exact dennis phenotypes are hard to determine relative to *H. melpomene*, but that do have orange patterns in both these wing regions, show a signal of shared ancestry with *H. melpomene aglaope* at only one of the two dennis modules; the larger 6kb window. Furthermore within this 6kb region, only one window is found to be shared and derived in both comparisons species (Figure 4.4c; Figure 4.4d). This window does not overlap with the previously defined dennis region, but is located directly next to it. It is possible that these two regions each control one of the two dennis phenotypes, the forewing and the hindwing dennis, with *H. pardalinus* having shared ancestry for one but not the other of these two phenotypes. Alternatively, it may be that only the smaller 3kb window that is shared between *H. elevatus* and *H. melpomene* races, but not between *H. melpomene* and *H. pardalinus* races controls the dennis, with *H. pardalinus* having a different 'dennis' phenotype.

The 3kb windows shared between *H. elevatus* and *H. melpomene* races, but not between *H. melpomene* and *H. pardalinus* races, appears to be derived from an ancestral silvaniform allele, suggesting that this dennis phenotype introgressed into *H. melpomene.* This was also the conclusion based on phylogenetic analysis in Wallbank *et al* (2016). However, the other 6kb window, like the pattern, seen for the wider *rays* locus, suggests independent evolutionary histories for each of the 3kb windows within it, with the window shared between *H. pardalinus butleri, H. elevatus* and *H. melpomene* races with the dennis, again derived from an ancestral Silvaniform allele and the other derived from an ancestral *H. melpomene* allele.

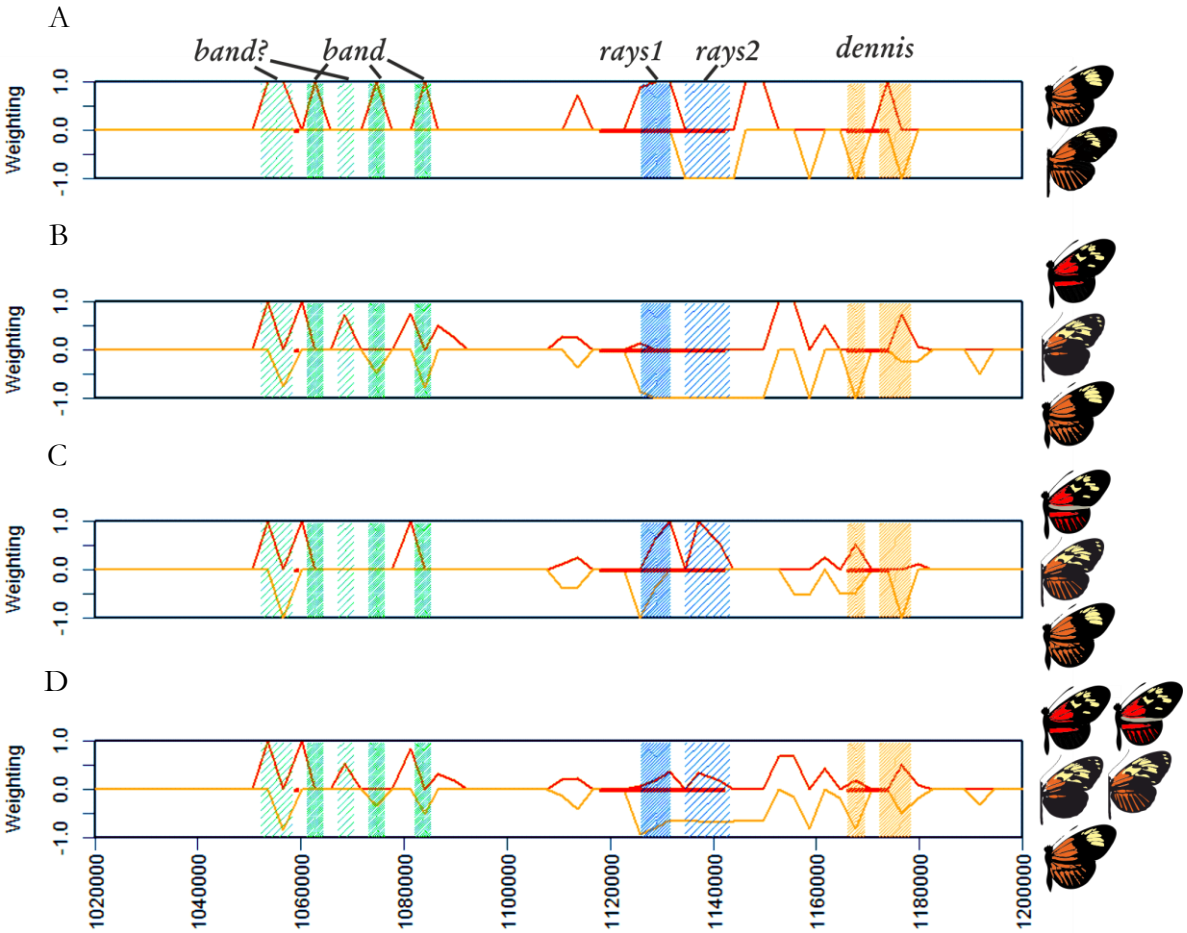


Figure 4.5 – Shared ancestry as topology weightings (proportions) around *optix* gene from phylogenetic comparisons. A) Shows ancestral source of loci shared (from Figure 4.4a) between *H. elevatus* and *H. melpomene aglaope*, negative from silvaniforms, positive from *H. melpomene*. B-D show phylogenetic comparisons that included both dennis-rayed (*H. elevatus bari and H. melpomene thelxiopeia*) and/or dennis only (*H. elevatus tumatumari* and *H. melpomene meriana*) Guianese taxa, along with *H. melpomene aglaope* from Peru. Taxa in order from top to bottom: B) *H. melpomene meriana*, *H. elevatus tumatumari* and *H. melpomene aglaope*; C) *H. melpomene thelxiopeia*, *H. elevatus bari* and *H. melpomene aglaope;* D) All dennis taxa, *H. melpomene thelxiopeia/meriana*, *H. elevatus tumatumari/ bari* and *H. melpomene aglaope*. Positive values show shared ancestry between middle taxa and top taxa, relative to bottom, negative values show shared ancestry between middle and bottom taxa, relative to top. Red shows from left to right, *optix*, and the previously defined *rays* and *dennis* regions. Shading: Green shows putative band modules, blue shows putative rays modules, orange shows putative dennis modules.

Topology weightings of comparisons across all Guianese taxa with the dennis along with *H. melpomene aglaope* from Peru, showed ancestry was mixed in *H. elevatus* (Figure 4.5d). With trees supporting both phylogenies that group Guianese forms of *H elevatus* and *H. melpomene* together, and phylogenies that group Guianese forms of *H elevatus* with the allopatric *H. melpomene aglaope*. This again suggests some level of gene flow may have occurred at these regions since the dennis alleles diverged from the *H. melpomene aglaope*/*H elevatus pseudocupidineus* alleles.

4.3.1.3 A putative band *locus*

As well as identifying narrower putative regulatory regions within the known *dennis* and *rays* modules, I identified a previously undefined region that showed shared ancestry between *H. elevatus* and co-mimics in *H. melpomene*. Although, the exact role of this region remains somewhat unclear, it appears that it may be involved in *optix* expression on the forewing in relation to the red forewing band. From crosses between *H. melpomene* races this is thought to be under epistatic control by loci that regulate both *optix* and *cortex*, with a number of intermediate wing patterns found in F2s (Sheppard *et al.* 1985). One of the difficulties in defining a *band* loci using association studies, and in this study, is uncoupling the dennis and rays phenotypes from that of the band phenotype, as no red-banded, rayed or dennis forms exists. While the experimental design here addresses this with regards to ray and band, using Guianese dennis-rayed *H. melpomene* and *H. elevatus* taxa and the dennis-only *H. melpomene*, which all share the same band phenotype. For the dennis and band phenotypes this is harder. This makes it possible that the putative *band* locus defined here is actually involved in dennis pigmentation. However, given the previous recombination breakpoint analysis that defined the approximate *dennis* region in Wallbank *et al* (2016), it seems likely that the region putatively identified here as *band* is correct.

In addition to the putative *dennis* and *rays* regions described in the previous sections, three 3kb windows were consistently seen to show shared ancestry between Peruvian and Guianese *H elevatus* and *H. melpomene* with yellow forewing bands, relative to both *H. pardalinus* (with Peruvian taxa) and postman forms of *H. melpomene* (Figure 4.4a; Figure 4.4b; Figure 4.4e; Figure 4.4f). Furthermore, *H. pardalinus races* show no signal of shared ancestry with either *H. melpomene aglaope* or *H. melpomene amaryllis* in this region, likely having alternative silvaniform alleles that regulates forewing *optix* expression (Figure 4.4c; Figure 4.4d). An additional 6kb window also showed shared ancestry between *H. melpomene aglaope* and *H elevatus pseudocupidineus* when compared to *H. pardalinus* (Figure 4.4a, light green shading), but in comparisons to *H. melpomene amaryllis* only one of these two windows showed shared ancestry (Figure 4.4b). This narrowed this region down. Furthermore, this 3kb window did not show shared ancestry between the Guianese *H. melpomene* and *H. elevatus* forms that have a yellow rather than red forewing band, relative to the postman form *H. melpomene melpomene*, while the other 3kb window did (Figure 4.4e; Figure 4.4f). These windows were therefore not consistently found, although a signal around this region was. Surprisingly, a single 3kb window in the same region, showed shared ancestry in the opposite direction, between *H. elevatus* and *H. melpomene* postman forms. While it is seems that this wider region is associated with *optix* expression on the forewing, any adaptive role of this region of opposite ancestry is unclear.

All of these windows of shared ancestry between *H. melpomene* and *H. elevatus* in this region, appear to have originated from ancestral *H. melpomene* alleles. This suggests that the lack of the red forewing band, is an ancestral *H. melpomene* phenotype (Figure 4.5a). Topology weightings from comparisons of Guianese taxa with yellow forewing bands, that also included *H. melpomene aglaope* from Peru, showed little evidence that gene flow may have occurred between these two species in the Guianas, since these alleles diverged from the *H. melpomene aglaope*/*H elevatus pseudocupidineus* alleles (Figure 4.5d).

4.3.2 **Regulatory control of *WntA* expression**

Between *H. melpomene*, *H. elevatus* and *H. pardalinus* there are three *WntA* phenotypes that differ in the melanic patterning seen across the discal portion of the forewing. These are, the complete black found in *H. melpomene aglaope*, *H. elevatus pseudocupidineus* and the *H. melpomene* postman forms, the silvaniform markings of *H. pardalinus butleri*, and the broken band forms of *H. elevatus* and *H. melpomene* in the Guianas. Pairwise phylogenetic comparisons using Peruvian taxa, identified regions showing shared ancestry around *WntA* between all races of *H. melpomene* and *H. elevatus*, relative to *H. pardalinus*. Furthermore across the rest of chromosome 10 there was no other strong signal of shared ancestry (Figure 4.6). However, the windows found to be shared derived across species, were not as great as that seen in the *BD* region on chromosome 18. This is due to the smaller number of shared regulatory regions around this gene, which is likely due to the smaller number of mimicry elements controlled by *WntA*. While shared ancestry, proposed to be the product of introgression between *H. elevatus pseudocupidineus* and *H. melpomene aglaope* has previously been identified around the colour pattern loci *BD* and *Yb* (Dasmahapatra *et al.* 2012), this is the first demonstration of the same effect around *WntA* (the *Ac*) locus.

Three separate regions around *WntA* showed a signal of shared ancestry between *H. elevatus pseudocupidineus* and *H. melpomene aglaope* and *H. melpomene amaryllis*, relative to *H. elevatus'* sympatric sister species *H. pardalinus butleri* and the allopatric race *H. melpomene rosina* (Figure 4.7a). These were a 6kb region downstream of *WntA*, a 3kb window found within a large non-protein coding region of the *WntA* gene, and a larger 9kb window up-stream of *WntA*, with one of these three windows overlapping with the very end of the *WntA* gene. All three of these loci appear to have been derived from ancestral *H. melpomene* alleles, suggesting that adaptive introgression into *H. elevatus* has led to a switch in colour pattern from a previously silvaniform type pattern (Figure 4.7b).

In the Guianas there were no regions of shared ancestry between the broken banded *H. elevatus tumatumari/bari* and *H. melpomene meriana/thelxiopeia*, relative to the postman form *H. melpomene melpomene* (Figure 4.7d). This suggests that the broken banded phenotype has evolved independently in each of these two species. This also meant that a putative regulatory region associated with this particular phenotype could not be identified in either. This independent origin for the broken banded phenotypes was supported by the lack of phylogenetic signal showing shared ancestry between Guianese taxa in the phylogenetic comparison including broken banded *H. elevatus* and *H. melpomene* from the Guianas, and *H. melpomene aglaope* from Peru (Figure 4.7f).

A phylogenetic signal of shared ancestry was seen between Guianese *H. elevatus* and both broken banded and postman *H. melpomene* in the Guianas, relative to the allopatric *H. melpomene rosina* and silvaniforms *H. ethila* and *H. hecale* (Figure 4.7e). This was at two of the putative regulatory regions identified in the phylogenetic comparisons using Peruvian taxa. This supports a role for these regions in *WntA* regulation, and suggests that the *H. elevatus tumatumari/bari* phenotype is a novel phenotype, derived from an ancestral *H. melpomene* allele. In the phylogenetic comparison including both broken banded sympatric comimics of *H. elevatus* and *H. melpomene* from the Guianas, and the allopatric *H. melpomene aglaope,* there was no strong phylogenetic signal of shared ancestry between the sympatric Guianese races relative to *H. melpomene aglaope* (Figure 4.7f). With trees instead grouping the *H. melopomene* taxa together and *H. elevatus* basal to these, this is consistent with a scenario in which the *H. elevatus tumatumari/bari* phenotype is derived from an ancestral *H. melpomene* allele, while gene flow persists amongst *H. melpomene* races.

Figure 4.6– Shared ancestry as Loess-smoothed topology weighting, across chromosome 10 in the pairwise comparison using *H. melpomene aglaope*, *H. elevatus pseudocupidineus* and *H. pardalinus butleri*. Positive values shows shared ancestry between comimics *H. melpomene aglaope* and *H. elevatus pseudocupidineus*, negative values shows shared ancestry between *H. melpomene aglaope* and *H. pardalinus butleri*. Red square shows the location of the *WntA* gene. Y-axis shows position in base pairs across the chromosome.

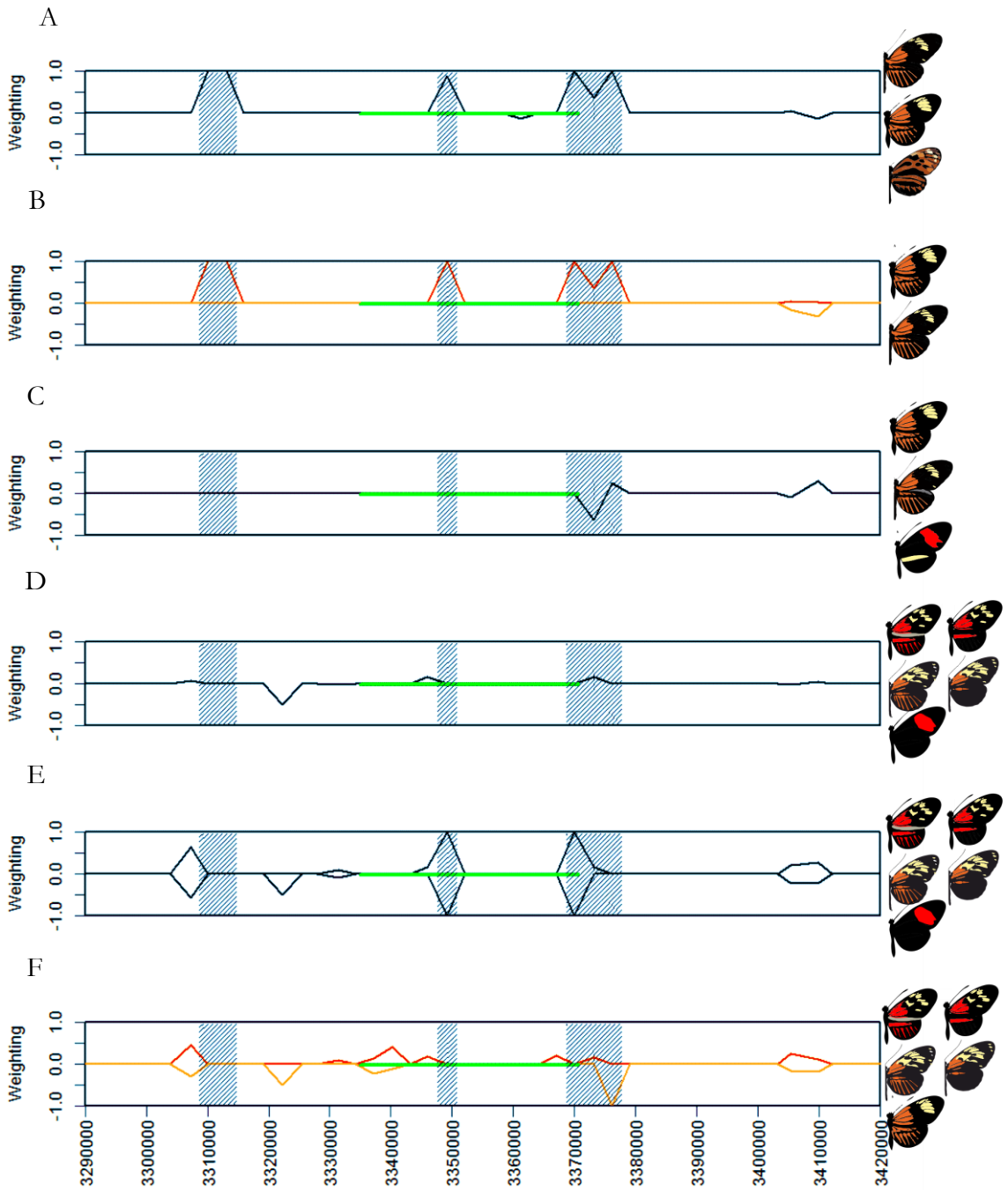


Figure 4.7 – Shared ancestry as topology weightings (proportions) around the *WntA* gene. A) shows pairwise phylogenetic comparison between *H. melpomene aglaope*, *H. elevatus pseudocupidineus* and *H. pardalinus butleri*; B) shows ancestral source of regions showing shared ancestry across species, C) pairwise phylogenetic comparison between *H. melpomene aglaope, H. elevatus pseudocupidineus* and *H. melpomene amaryllis*; D) pairwise phylogenetic comparison between *H. elevatus tumatumari/bari*, *H. melpomene meriana/thelxiopeia* and *H. melpomene melpomene*; E) separate topology weightings from comparison D, showing shared ancestry across all three taxa, F) phylogenetic comparison including *H. elevatus tumatumari/bari*, *H. melpomene meriana/thelxiopeia* and *H. melpomene aglaope*, showing possible shared ancestry with both. For A, C, D, E and F: +1 shows shared ancestry between middle and top taxa, -1 shows shared ancestry between middle and bottom taxa. For B, +1

(red) shows *H. melpomene* ancestral source, -1 (orange) shows silvaniform ancestral source. Green bars show WntA gene; shading shows putative *WntA* regulatory modules.

4.3.3 Regulatory control of *cortex* expression

Pairwise phylogenetic comparisons across both Guianese and Peruvian taxa, consistently identified a number of regions showing a shared and derived ancestry between taxa of different species with matching colour pattern phenotypes around the region that contains the gene *cortex*. Furthermore, this shared ancestry signal was far greater across this narrow region than across the rest of chromosome 15 in all comparisons (Figure 4.8). This is consistent with previous work that identified a signal of shared ancestry across co-mimics of different species in this region of adaptive importance (Dasmahapatra *et al.* 2012). A previous genome wide association study across *H. melpomene* clade taxa and species, identified two putative regions both associated with the yellow hindwing bar and yellow forewing band phenotypes, found at either ends of the gene *cortex* (Nadeau *et al.* 2016). The results presented here found are generally in agreement with this, with both these regions showing shared derived ancestry between taxa sharing phenotypes, relative to taxa that did not share these phenotypes (Figure 4.9).

In Nadeau *et al.* (2016) two putative regulatory regions around *cortex* were identified. These showed genotype by phenotype association with both the forewing yellow band and hindwing bar phenotypes. However, association with the yellow band was stronger in the upstream region, while the downstream region showed stronger associations with the yellow bar. In the analysis presented here, I identify a signal of shared ancestry across all pairwise comparisons between taxa sharing this yellow forewing band colour pattern phenotype, relative to those that do not, in this upstream, band region. A single 3kb window within this wider region, was consistently found to be shared and derived in all yellow banded *H. melpomene* and *H. elevatus* taxa, relative to postman forms (Figure 4.9b;

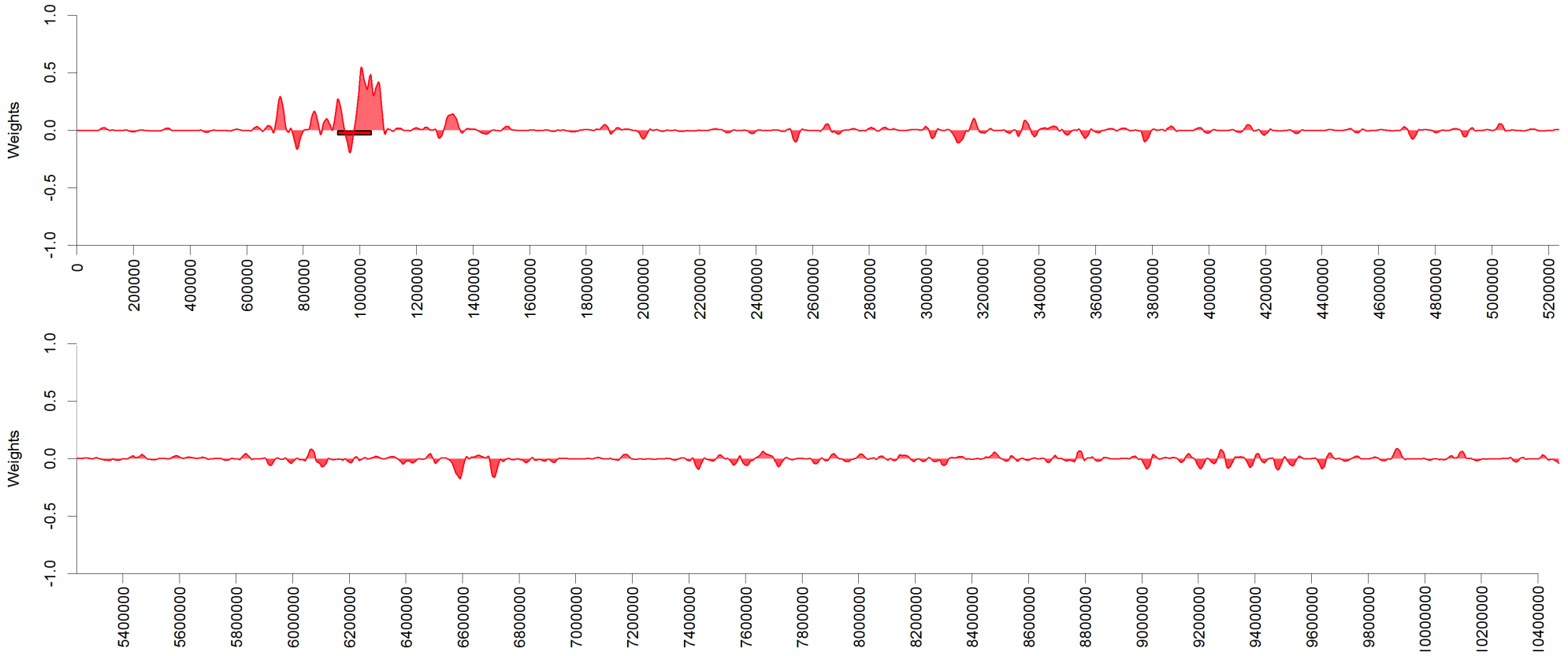


Figure 4.8 – Shared ancestry as Loess-smoothed topology weighting, across chromosome 15 in the pairwise comparison using *H. melpomene aglaope*, *H. elevatus pseudocupidineus* and *H. pardalinus butleri*. Positive values shows shared ancestry between comimics *H. melpomene aglaope* and *H. elevatus pseudocupidineus*, negative values shows shared ancestry between *H. melpomene aglaope* and *H. pardalinus butleri*. Red square shows the location of the gene cortex. Y-axis shows position in base pairs across the chromosome.

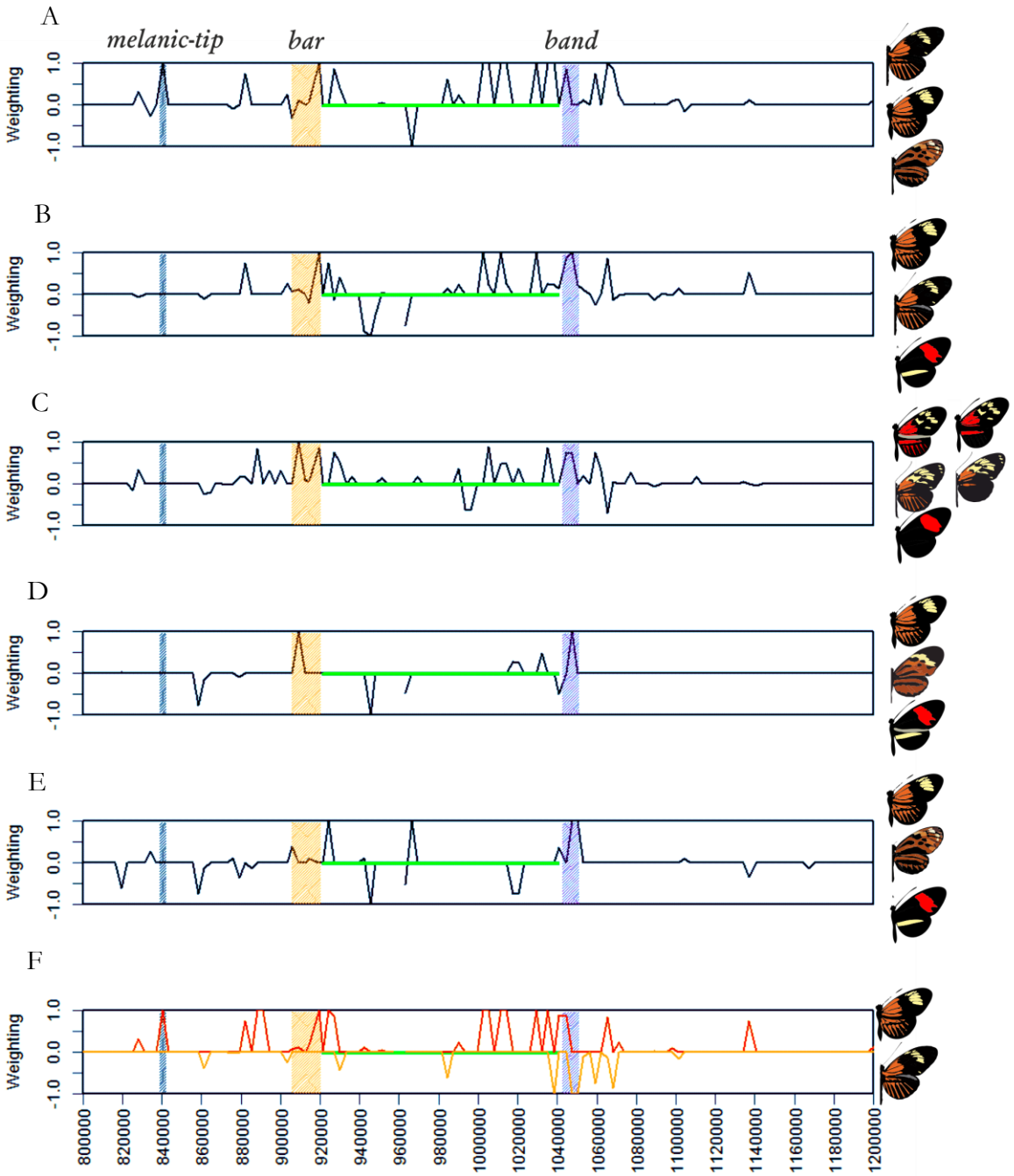


Figure 4.9 – A-E) Pairwise phylogenetic comparisons with shared derived ancestry as topology weightings (proportions) around the *cortex* gene. Primary taxa is in the middle and given here first, then top taxa, then bottom taxa. A) Comparison between *H. melpomene aglaope*, *H. elevatus pseudocupidineus*, *H. pardalinus butleri*; B) *H. elevatus pseudocupidineus*, *H. melpomene aglaope*, *H. melpomene amaryllis*; C) *H. elevatus bari/tumatumari*, *H. melpomene thelxiopeia/meriana*, *H. melpomene melpomene* D) *H. pardalinus butleri*, *H. melpomene aglaope*, *H. melpomene amaryllis*; E) *H. pardalinus sergestus*, *H. melpomene aglaope*, *H. melpomene amaryllis*. Positive values show shared ancestry between middle (primary) taxa and top taxa, relative to bottom, negative values show shared ancestry between middle (primary) and bottom taxa, relative to top. Green bars shows position of *cortex*. Shading: Blue, shows putative apical band/melanic forewing tip locus, orange, yellow hindwing bar; purple

shows putative band module. F) Shows ancestral source of loci shared between *H. elevatuas* and *H. melpomene aglaope*, negative from silvaniforms, positive from *H. melpomene*.

Figure 4.9d; Figure 4.9e). This is congruent with what was seen in Nadeau *et al* (2016). Furthermore, this 3kb window was also found to be shared derived between both races of *H. pardalinus* and *H. melpomene aglaope* (Figure 4.9d; Figure 4.9e), relative to *H. melpomene* postman forms. This was particularly clear in the comparison with *H. pardalinus sergestus*, in which this was one of only two windows shared with *H. melpomene aglaope* across the whole of chromosome 15 (Figure 4.9d). If this putative region, is the locus that regulates *cortex* expression for the yellow band, which *H. pardalinus sergestus* does appear to have, this means that *H. pardalinus butleri* also has this phenotype. This is quite plausible given the yellow pigmentation (mixed with orange) that can be clearly seen in this region of the forewing of *H. pardalinus butleri* specimens. This is also supported by the fact that windows around this region showed mixed ancestry, in particular this window was found to have silvaniform ancestry (Figure 4.9f), with *H. melpomene* taxa grouping amongst these taxa, rather than the other way around.

I also identify a signal of shared ancestry in the downstream bar region, in all pairwise comparisons between taxa lacking the hindwing bar phenotype, relative to those that have it, again supporting the finding that this region is involved in adaptive regulation of *cortex* (Figure 4.9b; Figure 4.9d; Figure 4.9e). However, there was no single window that was consistently found across taxa to be associated with this phenotype, and so the exact location of any regulatory region is unclear. Furthermore, regions of shared divergence were found between Guianese *H. melpomene* and *H. elevatus* lacking the yellow bar, relative to the postman form also lacking the bar (Figure 4.9c). Unfortunately, this further confounds the analysis, although it can be easily explained given the exchange of mimicry loci between *H. melpomene* and *H. elevatus* generally appears to have occurred in Peru rather than the Guianas, and the hindwing bar phenotype is found in postman races that *H. melpomene melpomene* is more closely related to.

The analysis of this hindwing bar phenotype is impaired by the lack of taxa of two different species that share the hindwing bar through likely introgression. An analysis of *H. timareta* and/or *H. beskei* and *H. melpomene* may therefore provide clearer results for this phenotype. What can be confirmed, is that the ancestral source of this region appears to be *H. melpomene* rather than silvaniform. Combining the finding of separate ancestries for each of these two previously identified regulatory regions with knowledge of the frequency of the two phenotypes, does support their respective roles, with the hindwing bar phenotypes from *H. melpomene*, where it is commonly found, and the yellow band from silvaniforms where similar phenotypes can easily be identified. It can however, not be ruled out that both regions play a role in patterning both phenotypes, what does seem apparent is that both these regions are of adaptive importance.

One more 3kb window was found across the cortex region was found to show a pattern of shared derivation congruent with differences in the melanisation of the forewing tip in *H. melpomene aglaope* and *H. elevatus pseudocupidineus* and the presence of the apical band in *H. pardalinus butleri.* Not only was it shared and derived in *H. melpomene aglaope* and *H. elevatus pseudocupidineus,* relative to *H. pardalinus butleri* (Figure 4.9a), but it was not shared derived in any other pairwise phylogenetic comparisons, where all taxa shared the same melanic tip. Furthermore, the ancestral source of this region was found to be from *H. melpomene*, further supporting this as a putative regulatory region, controlling one, or both of these phenotypes.

4.4 DISCUSSION

In this chapter I investigate patterns of shared ancestry around known colour pattern genes, between *H. melpomene* races and *H. pardalinus* and *H. elevatus* races that are associated with shared colour pattern phenotypes. I identify regions that are: 1)

consistently found to be shared between taxa from different species that have matching colour pattern phenotypes; 2) are derived in these taxa relative to close relatives that have different colour pattern phenotypes, and 3) are not shared between taxa of different species that do not have these matching colour pattern phenotypes. The confirmation of all three, suggests that a locus is divergent between related taxa that have different colour pattern phenotypes, but is shared between more distantly related taxa that share colour pattern phenotypes. This suggests the locus is under strong selection over hybrid zones and thus likely involved in colour pattern control (Baxter *et al.* 2010; Counterman *et al.* 2010; Nadeau *et al.* 2012).

I find elevated levels of this shared-derived signal, across three colour pattern loci (*BD*, *Yb* and *Ac*), across multiple comparisons, and relative to the background levels on other parts of these chromosomes, supporting a history of adaptive colour pattern introgression at all three colour pattern loci. I provide the first evidence of introgression around the gene *WntA*, which consistent with previous work that has suggested a history of introgression has shaped the colour patterns of these species (Dasmahapatra *et al.* 2012; Pardo-Diaz *et al.* 2012). Furthermore, the ancestral sources of these putative colour pattern loci are varied, with evidence suggesting loci moving across species boundaries in both directions between the *melpomene* and silvaniform clades. Not only was this signal of shared derived ancestry between taxa with matching colour patterns heavily concentrated around the three known colour pattern genes, but on a finer scale these shared derived regions matched a number of putatively defined regulatory modules identified through GWAS studies at both *cortex* and *optix* (Nadeau *et al.* 2016; Wallbank *et al.* 2016).

4.4.1 **A newly defined regulatory module of *optix***

At *optix* the analysis identified three narrow regions that were associated with colour pattern variation. Two of these were the already delimited *rays* and *dennis* regions (Wallbank *et al.* 2016). However, this analysis identified two separate windows of shared

ancestry within this previously identified *dennis* region, one that was shared derived between *H. melpomene* and *H. elevatus*, relative to *H. pardalinus* taxa, and another just next to the *dennis* region that was also shared between dennis *H. melpomene* and *H. pardalinus* taxa. I propose that these either both control dennis variation, perhaps one controlling the hindwing and the other the forewing dennis, or alternatively that the *dennis* region has been narrowed to just the 3kb window within the previously defined *dennis* module.

For rays, I also identified a region that corresponded to that found in previous work (Wallbank *et al.* 2016). This region in fact contains what appear to be two loci with distinct evolutionary histories, one ancestrally *H. melpomene*, and the other silvaniform. In comparisons between *H. melpomene* and *H. elevatus* races in the both Peru and the Guianas these both show variation with the presence and absence of the rays phenotype, assuming as appears most likely, that the non-rayed *H. elevatus* race appears to have lost the rays independently from a rayed allele. However, the silvaniform locus also showed shared derived ancestry between *H. pardalinus sergestus* and the *H. melpomene aglaope*. From the hindwing orange/black phenotypes in these two taxa, it is clear that the rays phenotype is created by simply increasing melanisation found around hindwing venation. Two hypotheses can explain this observed pattern, either only the *H. melpomene* derived locus controls the rays phenotype, in which case this analysis has putatively narrowed this *rays* region to a single 9kb locus, or alternatively the whole 21kb region is involved with each perhaps controlling a different aspect of the rays phenotype, and combining to produce the full effect.

Finally I identified a third region, much nearer to *optix* that appears to control either dennis or the forewing band phenotypes, as the signal for these two phenotypes could not fully be uncoupled. However, given the previous identification of the *dennis* locus, it seems more likely that I have identified a novel regulatory module controlling the

forewing band. This is also supported by the ancestral source of these loci, which appears to be *H. melpomene*. Furthermore, this module was not found to be shared between *H. pardalinus*, and yellow banded *H. melpomene* races. This is not unexpected, and suggests that three separate alleles exist within these taxa for forewing *optix* expression; a red banded *H. melpomene* allele, a non-red banded *H. melpomene* allele, and a silvaniform allele that also controls forewing Optix. *H. elevatus* appears to have acquired the *H. melpomene* haplotype, while *H. pardalinus* races have retained the silvaniform haplotype.

4.4.2 Conserved regulatory modules of *WntA*

I identified windows in two regions in and around *WntA* that showed a conserved signal of shared derived ancestry across comparisons with sympatric and parapatric *H. melpomene* and *H. elevatus,* relative to *H. pardalinus*, and the allopatric *H. melpomene rosina*. This is therefore a strong candidate for a region that controls black pigmentation across the discal portion of the forewing in *H. elevatus* and *H. melpomene*, that has introgressed between these two species. Furthermore the direction of this introgression appears to be from *H. melpomene* into *H. elevatus*, this signal again supports this hypothetical role for this module, given the likely ancestral phenotypes in each taxa. This is the first time a signal of introgression at *WntA* has been identified between any *Heliconius* species.

4.4.3 Regulatory modules of *cortex*

We identified two major regions showing shared derived ancestry between taxa of different species sharing phenotypes, relative to taxa that did not share these phenotypes. However, the signal compared to that around *optix*, was both harder to interpret and less consistent. This analysis was particularly hampered by lacking a comparison in which the yellow bar was the shared derived phenotype. However, on a broader scale the analysis was consistent with previous GWAS analysis, across *H. melpomene* clade taxa and species, that identified the same two putative regions I identified here. Both of these were associated with the yellow hindwing bar and yellow forewing band phenotypes, found at

either ends of the gene *cortex* (Nadeau *et al.* 2016). This could be explained both by noise which is feature of GWAS studies, especially those over multiple species, and the difficulty of uncoupling these phenotypes, or alternatively because both regions do contribute to the control of both phenotypes.

Given the taxa available, uncoupling variation due to these two phenotypes, as well as forewing tip melanisation and the apical band was difficult in this study. Therefore, any conclusions that can be drawn are perhaps less clear and more speculative than those drawn for *optix* or *WntA*. However, across comparisons the analysis did identify the two regions previously found from GWAS analysis (Nadeau *et al.* 2016). Furthermore, under the assumption that the yellow forewing band phenotype is present in *H. pardalinus butleri*, I also identified a single 3kb window that showed perfect association with the yellow forewing band. This assumption is not unreasonable given *H. pardalinus butleri* does have yellow pigmentation in this part of the forewing, not dissimilar from that seen in *H. melpomene* and *H. elevatus* taxa. This window, as well as others around it, was found to come from a silvaniform ancestral source, suggesting that the direction of introgression of this putative *band* locus may be from silvaniforms into *H. melpomene*. This window was also one of only two across all of chromosome 15, at which all trees showed shared derived ancestry between *H. pardalinus sergestus* and *H. melpomene aglaope*, relative to *H. melpomene amaryllis*, with the other window at the putatively defined regulatory region that shows strongest association with the yellow hindwing bar. Furthermore, this yellow hindwing bar associated region had a *H. melpomene* ancestral source, which also supports the idea that this may be the locus that controls the yellow hindwing bar. Together this makes this second window, a weaker candidate for the band phenotype, than the window found at the putative *cortex band* locus.

In addition, to these loci, I also identify a single 3kb region that designate as a putative regulatory region perhaps involved in the control of either forewing tip melanisation

and/or the apical band. This locus was found to have a *H. melpomene* origin as did other windows of shared derived ancestry around this region, identified in other comparisons, supporting the idea it controls forewing tip melanisation, which is seen in all *H. melpomene* races but not in many silvaniforms.

4.4.4 **Modulation of mimicry and pattern switching**

The findings from this analysis are concordant with findings from previous GWAS studies that had identified some of these regions as regulatory modules (Nadeau *et al.* 2016; Wallbank *et al.* 2016). However, here I have been able to identify several more putative modules associated with pattern variation, as well as narrowing known modules. My results therefore indicate that this *cis*-regulatory modulation of enhancers is common across mimicry genes. This modulation provides a flexible toolkit through which gene expression changes can rapidly alter phenotypes and drive adaptive evolution (Wray 2007). This modulation is frequently seen in cases of parallel and convergent evolution, as deletion or mutation at a single enhancer is enough to have a major phenotypic effect (Chan *et al.* 2010; Frankel *et al.* 2012). However, in *Heliconius* it appears to also facilitate adaptive evolution through the swapping of these enhancers between lineages and species, without otherwise having major detrimental fitness effects (Wallbank *et al.* 2016). While this has already been shown for *dennis*, and for part of the *H. melpomene rays* locus (Wallbank *et al.* 2016), I find evidence of this between *H. elevatus* and *H. melpomene* for other additional regulatory modules as well as at *WntA*. Furthermore this can be expanded to other taxa, in this case *H. pardalinus sergestus*, which also shows a signal of adaptive introgression with *H. melpomene* at two putatively identified cis-regulatory modules around *cortex*, and at a locus that may form part of the *rays* module. Furthermore, the evidence in fact suggests that two of these modules are in fact derived from a silvaniform ancestral state. This indicates porous species boundaries in the *H. melpomene*/silvaniform clade that has led to frequent adaptive pattern shifts across taxa

through enhancer shuffling, which has shaped adaptive evolution across both silvaniform and *H. melpomene* taxa.

4.4.5 **Independent mimicry in the Guianas**

Given this promiscuous exchange of adaptive colour pattern loci observed between Peruvian taxa, it was striking that this pattern was not seen between mimetic races *H. melpomene meriana* and *H. elevatus tumatumari* in the Guianas. As expected, given the presence of a postman race in the Guianas, the rays appears to have been lost in the dennis only *H. melpomene meriana* through recombination with this postman form. This is supported by both the analysis here, and from previous analyses as well (Wallbank *et al.* 2016). In contrast however, the *rays* allele in non-rayed Guianese *H. elevatus tumatumari* appears to be more similar to that found in rayed Peruvian *H. elevatus pseudocupidineus/aglaope*. This suggests the rays phenotype has been lost independently in *H. elevatus*, despite the possible opportunity for this to have occurred via introgression, which is suggested by the signal of shared derived ancestry between the rayed *H. elevatus* and *H. melpomene* forms in the Guianas relative to *H. melpomene aglaope*. Furthermore, the broken band phenotype in the Guianas, also appears to have evolved independently in both *H. elevatus* and *H. melpomene*, no shared derived regulatory modules found between them. Given this independent evolution, it seems possible that this broken banded phenotype in *H. elevatus* may have either been ancestral, or accrued through recombination with some other silvaniform taxa, as somewhat similar patterns exist in Ecuadorian *H. pardalinus* races, and as an F2 phenotype in crosses between *H. elevatus* and *H. pardalinus* (*unpublished data*). It would be interesting to assess the timings of these various introgression events, and the relationships across a broader range of taxa, in order to arrive at biogeographic hypothesis of how mimicry in these taxa has evolved, this could be achieved through dated phylogenies at some of these putative regulatory regions.

4.4.6 **Conclusion**

In this chapter I use a combination of the natural colour pattern diversity from two *H. melpomene* hybrid zones and a history of shared ancestry between these taxa, and *H. elevatus* and its sister species *H. pardalinus*, to identify and narrow putative regulatory modules around three major mimicry genes. This signal of shared ancestry also suggests that the ancestral sources of these putative colour pattern loci are varied. This indicates that at times species boundaries in the *H. melpomene*/silvaniform clade have been porous allowing frequent pattern shifts across taxa via adaptive introgression, shaping the evolution of mimicry across both silvaniform and *H. melpomene* taxa. However, between *H. elevatus* and *H. melpomene* in the Guianas the signal of shared ancestry suggests the independent evolution of mimicry, of both the broken band and non-rays phenotypes.

5. Pre-zygotic barriers between two sister species

5.1 INTRODUCTION

Although the theory of sympatric speciation has been around for well over a century (Darwin 1859), for much of this time it has been largely overlooked by zoologists. This is because gene flow and recombination make divergence difficult in sympatry, as they work together to destroy any linkage between traits that might otherwise characterise newly evolving species (Mayr 1963; Felsenstein 1981). For many years, this theoretical objection led to most speciation research focussing largely on the role that selection and mutation play in driving divergence between allopatric populations (Schluter 2009). More recently this focus has begun to change due to the availability of new molecular phylogenetic datasets that allow the empirical testing of suspected cases of sympatric speciation (Meyer *et al.* 1990; Savolainen *et al.* 2006; Geiger *et al.* 2010). This has occurred in concert with a shift away from defining speciation along geographical lines, where sympatry was speciation 'without geographical isolation' and allopatry the opposite (Mayr 1963). Now the vast majority of speciation is thought to occur at some point between these two extreme ends of the gene flow continuum (Bolnick & Fitzpatrick 2007; Fitzpatrick *et al.* 2008; Mallet *et al.* 2009).

In their seminal book 'Speciation' Coyne & Orr (2004) argued that to convincingly prove that speciation has occurred in sympatry the example must stand up to a number of criteria. These were as follows: i) species arising in sympatry should have overlapping ranges; ii) speciation should be complete; iii) species should be sister species or at least

monophyletic clades; iv) existence of an allopatric phase must be unlikely. This last step has been argued to be one of the reasons that sympatry is hard to prove, as finding examples where one can conclusively rule out any allopatric phase is very difficult (Bolnick & Fitzpatrick 2007). However, the breakdown of the dichotomy between allopatry and sympatry has led to the emergence of an increasingly large number of examples of speciation where gene flow has certainly occurred during divergence, some of which may have occurred in complete sympatry. These include Cameroonian crater lake cichlids (Schliewen & Klee 2004; Martin *et al.* 2015a), Nicaraguan crater lake cichlids (Barluenga *et al.* 2006), the tunicate *Ciona intestinalis* (Roux *et al.* 2013), as well as several examples in *Heliconius* butterflies (Salazar *et al.*, 2005; Martin *et al.*, 2015).

This shift in how gene flow is perceived has led to a focus on the mechanisms driving speciation when there are varying levels of gene flow between the diverging taxa. Otherwise ignoring how divergence can occur in the face of the homogenising effects of gene flow and hybridization, would overlook much of the complexity of speciation. One model of how speciation proceeds in the face of on-going gene flow is the 'Islands of divergence' hypothesis (Wu 2001). At the centre of this hypothesis is the idea that differences in just a few key traits can lead to reproductive isolation. In this scenario, if one was to look at divergence across the genome, strong divergent selection at regions of the genome controlling these speciation traits would look like ‘islands of divergence’ in a sea made up of an otherwise homogenous genome, where gene flow can occur freely (Nosil *et al.* 2009). This is important as it means that divergence can occur in the face of gene flow so long as selection at genomic regions controlling speciation traits is strong enough to overcome it.

It is hypothesised that the types of traits likely to be under the control of genes within islands of divergence are those directly involved in the processes of mate choice or resource use (Wu 2001), whether that is the mate recognition systems of *Drosophila* (Wu

et al. 1995), loci controlling growth differences between sympatric dwarf (limnetic) and normal (benthic) whitefish ecotypes (Rogers & Bernatchez 2004), or the wing colour patterns of *Heliconius* butterflies. In *Heliconius* these bright colour patterns are one of the best systems for testing the 'islands of divergence' model, as they act both as aposematic signal in Müllerian mimicry, and have been found to be used as cues for mate choice and species recognition, making colour pattern a so called 'magic trait' (Jiggins *et al.* 2001; Merrill *et al.* 2011). Colour pattern has also been found to be controlled by a relatively small number of loci spread across the genome. Furthermore, across a narrow hybrid zone between divergent subspecies of the species *H. melpomene* (Baxter *et al.* 2010) and *H. erato* (Counterman *et al.* 2010), the relevant colour pattern loci were found to be true islands of divergence, with divergence high in these loci but otherwise low across the rest of the genome (Dasmahapatra *et al.* 2012; Nadeau *et al.* 2014).

Between the sister species *H. pardalinus* and *H. elevatus* there lies the perfect opportunity to test the 'islands of divergence' hypothesis within the framework of speciation. *H. elevatus* colour pattern variation corresponds to variation in its Müllerian co-mimic *H. melpomene* as part of the dennis-rayed mimicry ring, while its sister species *H. pardalinus* is part of the silvaniform mimicry ring, which includes Ithommine butterflies (Brown 1976). In Peru, where *H. elevatus pseudocupidineus* and *H. pardalinus butleri* are sympatric, evidence of introgression of adaptive colour pattern loci, has been found between *H. elevatus* and its comimic *H. melpomene.* This evidence is in the signal of phylogenetic discordance at colour pattern loci, explored in chapter 4 (Dasmahapatra *et al.* 2012; Wallbank *et al.* 2016). However, low divergence across most of the genomes of *H. elevatus* and *H. pardalinus* is explained by extensive gene flow between the two species, at loci not associated to colour pattern. Therefore divergence must have occurred with some gene-flow, most likely through one of just two scenarios. In the first speciation occurs in sympatry with gene flow throughout divergence, while the second includes a

phase of allopatry without gene flow, followed by secondary contact and the resumption of gene flow.

It has been hypothesised that the introgression between *H. elevatus pseudocupidineus* and *H. melpomene* may have caused the divergence of *H. elevatus* and *H. pardalinus* (Dasmahapatra *et al.* 2012; Wallbank *et al.* 2016). If this introgression of colour pattern genes between *H. melpomene* and *H. elevatus* was the cause of speciation between *H. elevatus* and *H. pardalinus,* then speciation could have occurred in sympatry with gene flow throughout divergence, as colour pattern is both an ecologically important trait, and one with a secondary role in mate recognition and sexual selection. This dual selective role is necessary for sympatric speciation, as sexual selection on traits involved in mate choice and recognition alone are not able to drive sympatric speciation (Arnegard & Kondrashov 2004). If sympatric speciation did occur, then colour pattern preference is expected to be a strong reproductive barrier between these two species, with other barriers not as important as they will have arisen secondarily.

If speciation occurred with a phase of allopatry, then all possible reproductive isolation barriers are just as likely to have arisen first, and any single one could be as strong as any other. Although differences in colour patterns and preferences are hypothesised to have played a major role in the diversification of *Heliconius* (Jiggins 2008), other prezygotic barriers are likely to have also been important. Pheromones are well known in many Lepidoptera to play an important role in finding and attracting a mate. Differences in pheromone composition are thus another potential barrier to gene flow between species, which can lead to reproductive isolation. Most research into the role of pheromones in Lepidoptera have thus far focussed on moths (Lofstedt 1993; Symonds & Elgar 2008). Broadly speaking pheromones can be split into two classes; long-range and close-range signals. It is this second class, that have a role in courtship behaviour (Hartlieb & Anderson 1999), and that have been found to play an important role the mating systems

of butterflies like *Bicyclus anynana* (Costanzo & Monteiro 2007) and *Pieris napi* (Andersson *et al.* 2007). Although currently the role of this class of pheromones in *Heliconius* courtship has yet to be tested, it is likely they play an important role, as *Heliconius* courtship often consists of a long hovering stage during which it is hypothesised the male emits his pheromones (Klein & de Araújo 2010). In addition, without short range pheromones it is hard to otherwise explain how so many *Heliconius* species often coexist sympatrically and yet share the same colour pattern, and yet in other cases species can have multiple colour pattern races that so freely hybridise (Jiggins 2008). It is clear therefore that traits other than colour pattern, such as short range male sex pheromones, must play an important role in mate choice in *Heliconius*.

Colour pattern loci are known to be 'islands of divergence' between *H. elevatus pseudocupidineus* and *H. pardalinus butleri*. I therefore first tested the hypothesis that 1) there were significant colour pattern preference differences between males of the two species; and 2) the role that pheromones play in reproductive isolation between *H. elevatus* and *H. pardalinus*. If sympatric speciation occurred, then colour pattern preference is expected to be a strong reproductive barrier between these two species. In contrast, other barriers will be less important as they will have arisen after colour pattern has already largely reproductively isolated the species. Pheromones were investigated by testing the hypothesis that there were significant differences in the composition of chemical extracts from male androconial regions (shown in Figure 5.1) which are known in Lepidoptera to be the site of scent glands used to emit pheromones during courtship (Costanzo & Monteiro 2007). This sampling design, using chemical extracts from *H. elevatus pseudocupidineus* and two colour pattern races of *H. pardalinus*; *H. pardalinus butleri* and *H. pardalinus sergestus*, also allows comparisons to be made between allopatric taxa (*H. elevatus pseudocupidineus* vs *H. pardalinus sergestus*), parapatric taxa (*H. pardalinus sergestus* vs *H. pardalinus butleri*) and sympatric taxa (*H. elevatus pseudocupidineus* vs *H. pardalinus butleri*). If as hypothesised pheromones play an important role in reproductive

isolation, the degree of difference in pheromone composition should be greatest between sympatric taxa (*H. elevatus pseudocupidineus* vs *H. pardalinus butleri*) and least between allopatric taxa (*H. elevatus pseudocupidineus* vs *H. pardalinus sergestus*) where pre-zygotic barriers are not required. Together this work takes the first step towards confirming the importance of pheromones and colour pattern in the speciation of *H. elevatus* and *H. pardalinus,* as well as investigating whether the divergence of *H. elevatus* and *H. pardalinus* occurred in sympatry without a phase of reduced gene flow, or with gene flow but with a phase of allopatry.

5.2 METHODS

5.2.1 **Colour pattern preference**

In order to test the hypothesis that there are significant colour pattern preference differences between males of *H. pardalinus* and *H. elevatus,* male mate choice experiments were carried out. These were conducted in a 1.5m (L) x 1.5m (W) x 2m (H) cage using models made from dissected female wings. Pheromones were removed by washing the wings with the solvent dichloromethane. These models were placed 58cm apart from each other at a height of approximately 1.25m. Models were attached to cable ties, and manipulated in such a way as to simulate flight. In each 25 minute observational period four male *H. pardalinus* and four male *H. elevatus* were simultaneously presented with the model butterflies of each species, and in this time a number of male responses were recorded. These responses were approach of a model (defined as clear, directed flight to within 10cm of a model), and two courtship responses: alightment and hovering. This last behaviour is a stereotypical courtship behaviour defined by Klein & de Araújo (2010) as a male remaining in 'flight over the alighted female (5–15 cm) without considerable displacement'. A male was adjudged to have been active in an observational period if he exhibited any of these behaviours towards a model in the 25 minutes. Males were tested

in up to a maximum of three active observational periods. From this data, both courtship and approach probabilities for each species were calculated using the following maximum likelihood model (McMillan *et al.* 1997):

$$\ln(L) = \sum (\pi_i \ln (P_j) + E_i \ln (1 - P_j))$$

where πi is the total number of courtship events by male i directed towards *H. pardalinus* model, E_i is the total number of courtship events by male i directed towards *H. elevatus* model and P_j is the probability of males of species j performing behaviour directed towards *H. pardalinus*. Support limits equivalent to 95 per cent confidence intervals were obtained by searching for values that decreased the ln(L) by two units (Merrill *et al.* 2011).

A binomial generalised linear mixed effect model using the package lme4 1.1-12 (Bates *et al.* 2015) in R v3.1.2 (R Core Team, 2014) (GLMM), was used to evaluate courtship preference, with species used as the fixed effect, while trial and individual were used as random effects. Using trial as a random effect controls for a number of possible biases, such as time of day, temperature and the influence of the other individuals in the trial. Likelihood ratio tests (LRT) using the Stats package in R v3.1.2 (R Core Team, 2014), were then used to compare this model to a null model where species was not a factor, in order to test whether there was a significant difference in the strength of preference for conspecifics in each species. This null model was then compared to a model in which preferences were forced to be random, in order to test whether preference for conspecific butterflies was significantly different from this random model.

5.2.2 **Pheromone GC-MS**

In butterflies male sex pheromones are usually produced in the scent glands, and emitted from differentiated scales on the hindwings called androconia (Rutowski 1980), which are

clearly visible in *Heliconius* as grey/brown patches on the dorsal or the ventral part of male wings. Androconial and control (non-androconia) regions (see Figure 5.1) of wings were removed using tweezers and scissors that had been rinsed in dichloromethane and then allowed to dry. Wing tissue was then placed in 300 µl of dichloromethane in a 1.5ml glass vial. For each butterfly, control regions were sampled first so as not to cross contaminate from androconial regions. This process was carried out for five approximately 21 day old males of each *H. elevatus pseudocupidineus* and *H. pardalinus butleri* and *H. pardalinus sergestus* from captive stocks in Peru (by Lucie Queste). These control regions were used to get a baseline of chemical composition and quantity from across the wings, so as to isolate compounds specific to the androconia. One control from a *H. elevatus pseudocupidineus* individual showed clear signs of contamination leaving four controls for this species. Additional controls were also taken from regions corresponding to the androconia of two approximately 21 day aged-matched females of each H. *elevatus pseudocupidineus* and *H. pardalinus butleri.* These controls identify compounds that are found in both males and females, and therefore are not used as male sex pheromones. Sampled individuals were aged matched in order to control for variation due to age, and to ensure that all males were sexually mature, as male have been found not to mate until several weeks after eclosion (Jiggins and Mallet pers. comm.) Gas chromatography coupled to mass spectrometry (GC-MS) was used to analyse these extracts (by Florian Mann at the Technische Universität Braunschweig). An internal standard of tridecyl acetate was used so the amount (nmoles) of each compound, in each sample could be calculated. For full details of the GC-MS protocol, refer to Vanjari *et al* (2015). An additional ten captive bred male individuals of each *H. elevatus pseudocupidineus* and *H. pardalinus butleri*, were also sampled from populations in York (by myself), unfortunately for these samples the internal standard failed, this meant that only relative abundance of each compound could be calculated for these samples, rather than nmole amounts.

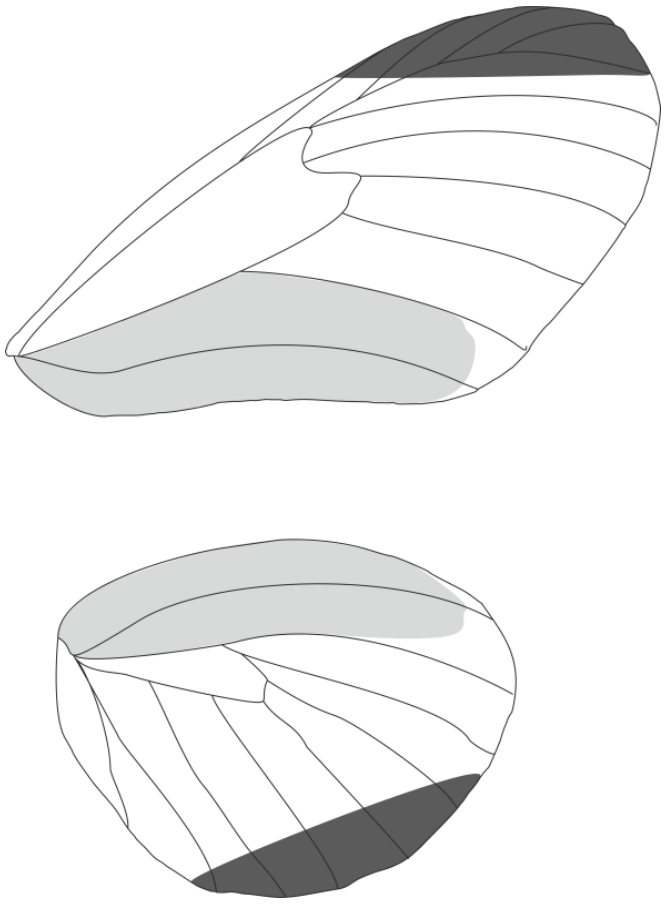


Figure 5.1 - Diagram showing in light grey the position of the forewing and hindwing androconia, and in dark grey are non-androconial controls regions.

The 'full compound dataset' contained the amount (nmoles) in all extracts from Peruvian samples (male androconial, male controls, female 'androconial'), of every compound found in at least one male androconial extract. Kruskall-Wallis tests were used in R v3.1.2 (R Core Team, 2014), to test for significant differences in the mean amount (nmoles), and number or compounds found between groups (male androconial, male controls, female 'androconial'). Nemenyi post-hoc tests from the R library PMCMR (Pohlert 2014) were then used for pairwise multiple comparisons (Tukey distribution). A principal component analysis was then carried out in R v3.1.2 using the *prcomp* function from the Vegan package (Oksanen *et al*, 2015) on this 'full compound dataset', with data centred and scaled, to investigate differences between species, wing regions, and sexes. Following this, Kruskall-Wallis tests were carried out on the 'full compound dataset' to identify candidate pheromone compounds. Those compounds found to show significantly different

abundances in extracts from male androconia than in both male and female control extracts were retained to produce a 'candidate pheromone dataset' of male androconial compounds only. Kruskall-Wallis tests were subsequently carried out on this 'candidate pheromone dataset' to see if they showed significant variation between species. Compounds that showed significant variation among species were retained to produce a 'species pheromone difference dataset'. Again principal component analysis was then carried out with data from the 'species pheromone difference dataset' centred and scaled.

To test for significant differences between taxa, pairwise Euclidean distances were then calculated between each sample. Mantel tests were then carried out using the ade4 package (Dray & Dufour 2007) in R v3.1.2 to compare these to simulated distance matrices. To first test whether there were significant differences between all taxa, a simulated matrix was used in which distances between samples from different taxa, were higher than those between samples of the same taxa. Further Mantel tests were then used to determine whether *H. elevatus* was more distinct from sympatric *H. pardalinus butleri* in its pheromone composition, than to the allopatric *H. pardalinus sergestus*. In these matrices, distances between samples of the same species were set to 0.0001. Then distances were varied so they were greater between *H. elevatus* and *H. pardalinus butleri* than those between *H. elevatus* and *H. pardalinus sergestus*, or those between *H. pardalinus butleri* and *H. pardalinus sergestus.* In other matrices the distance was set to be greater between *H. elevatus* and *H. pardalinus sergesus*, than in the other comparisons. The strength of the covariance between Euclidean distances from the PCA and simulated matrices could then be compared, to assess whether *H. elevatus* was more distinct from sympatric *H. pardalinus butleri* in its pheromone composition, than to the allopatric *H. pardalinus sergestus*

A principal component analysis was also carried out including both those samples already used above from the captive stocks in Peru, as well as the York samples. As the internal

standard had failed for the York samples, nmole amounts were converted to relative abundances (these relative abundances were over all compounds identified in the 'full compound dataset'). Kruskall-Wallis tests were carried out on these relative abundances for all compounds that were in the 'candidate pheromone dataset' (i.e had shown significant different in amount between male androconia relative to controls,) to see if compounds showed significant variation between species. Those compounds that did were retained in a new dataset termed the 'York sample abundance dataset'. Again, principal component analysis was carried out on this dataset in R v3.1.2 using the *prcomp* function (Oksanen *et al*, 2015) with data centred and scaled.

5.3 RESULTS

5.3.1 **Colour pattern preference**

A total of 147 approaches were recorded for 35 *H. pardalinus*, showing a slight 0.557 preference for conspecific models. Fewer approaches were recorded for *H. elevatus* with just 68 approaches from 24 males, but a higher proportion of these 0.691 were towards conspecific models (Figure 5.2). Where courtship is defined as hovering or alightment, a total of 97 courtships were recorded from 29 different *H. pardalinus* males, showing preferential courting of conspecific models with a probability of 0.628. Male *H. elevatus* were less responsive and more selective than *H. pardalinus* males, with 27 of 37 courtships from a total of 19 males towards conspecific models, giving a conspecific courtship probability of 0.729 (Figure 5.2). Furthermore, this preference for courting conspecifics was found to be significantly different from random with males showing a significant preference for courting conspecifics over heterospecifics (LRT, $\chi^2 = 12.743$, P = 0.0003).

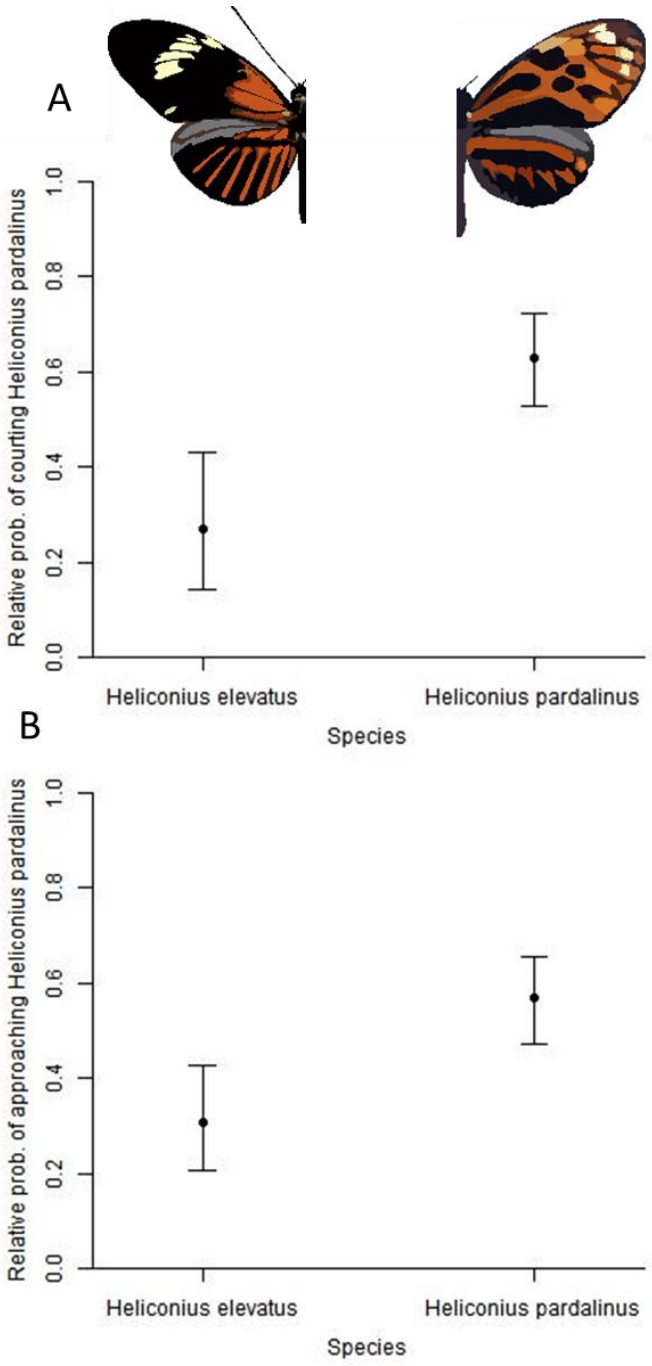


Figure 5.2 – A) shows the probability of male *Heliconius pardalinus* and *Heliconius elevatus*, courting models of *Heliconius pardalinus* in choice experiments. B) shows the probability of male *Heliconius pardalinus* and *Heliconius elevatus*, approaching models of *Heliconius pardalinus* in choice experiments. Error bars were obtained by searching for values that decreased in ln(L) by two units, these are equivalent to 95% confidence intervals.

5.3.2 **Pheromone composition**

GC-MS analysis of all extracts produced a 'full compound dataset' of 57 compounds, each found in at least one male androconial extract. From this dataset, extracts from the male androconia of all species were found to contain significantly greater mean amounts of compounds than male controls ($P < 0.001$) and female androconia ($P = 0.011$). In addition, male androconia of all species were found to contain significantly were also found to have a greater number of compounds than male controls ($P = 0.034$) and female androconia ($P < 0.001$) (Table 5.1 and Figure 5.3). No significant difference was found between the mean amount of compounds in male controls and female androconia ($P = 0.983$) or in the number of compounds found ($P < 0.999$). This was also supported by the principal component analysis of this 'full compound dataset' (Figure 5.3). This analysis suggests that as in other Lepidoptera, the androconial regions are the site of emission for male sex pheromones in *Heliconius*. Of these compounds, 28 were found by Kruskall-Wallis tests to show significant variation between extracts from male androconia regions and controls these were therefore deemed to be 'candidate pheromones'. Twenty of these 'candidate pheromones' were found by Kruskall-Wallis tests to show significant variation between the species (see Appendices 5, 6 and 7 for Kruskall-Wallis results).

	Male Andro.	**Male Control**	**Female Control**
H. elevatus	20.75±3.66 nmol 23.40	0.77±0.42 nmol 12.80	0.96±0.01 nmol 14.5
H. p. butleri	34.19±14.12 nmol 32.20	2.61±1.30 nmol 17.50	2.42±0.52 nmol 15.5
H. p. sergestus	9.63±3.50 nmol 26.40	1.33±0.50 nmol 14.40	No samples analysed

Table 5.1 - Values above show mean amount (± standard deviation) of total compounds found in extracts from male androconia, male controls and female controls, of all three species. Values below indicate mean number of compounds found.

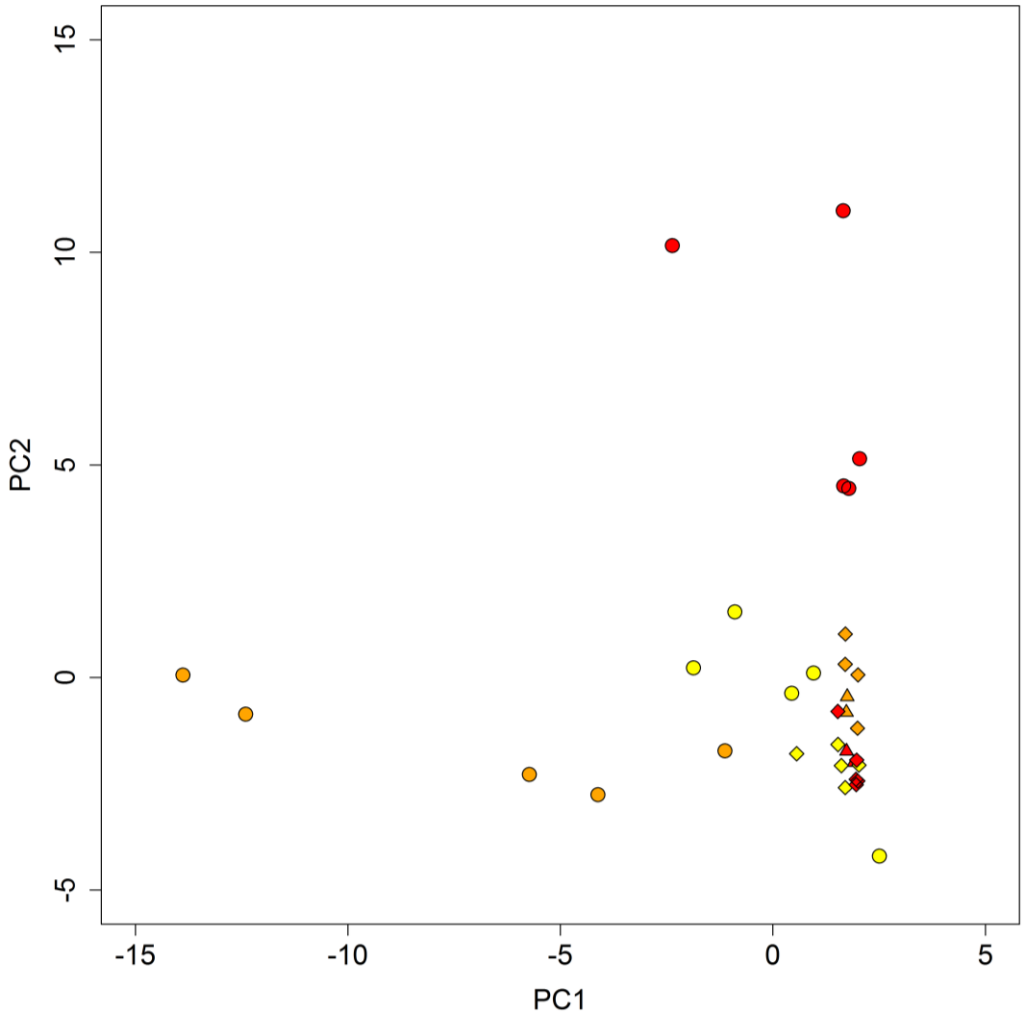


Figure 5.3 - Principal component scores from analysis on the 'full compound dataset'. Shape key: circle, male androconia; diamond, male control; triangle, females. Colours: orange, *H. pardalinus butleri*; yellow, *H. pardalinus sergestus*; red, *H. elevatus*.

Principal component analysis on this 'species pheromone difference dataset' produced two large Principal components PC1 and PC2, which explained 53.69% and 15.95% of variation respectively. PC1 discriminated between all three taxa while PC2 discriminated between *H. pardalinus sergestus* and the other two taxa (Figure 5.4A). Other Principal components did not describe variation between the species and so were not of interest to this study. Plotting variable loadings shows that there are two main clusters of correlated variables that contributed to PC1 and PC2 (Figure 5.4B).

Mantel tests (Table 5.2) found that pairwise Euclidean distances from PC1 and PC2 were found to be more significantly different between species than within species (Mantel test, r = 0.674; P = 0.0001). Significant covariance was found between pairwise Euclidean distances and all simulated matrices. However, the covariance between Euclidean distances from the PCA and the simulated data was strongest when the distance between *H. elevatus* and *H. pardalinus butleri*, was two times greater than that between other comparisons (Mantel test, r = 0.799; P = 0.0001). Significant covariance was also found between matrices in which distances were greater between *H. elevatus* and *H. pardalinus sergestus*. However, the covariance between Euclidean distances from the PCA and the simulated data was strongest when the distance between *H. elevatus* and *H. pardalinus sergestus* was just 0.5 times greater (Mantel test,r = 0.521; P = 0.0003). From these results it was clear that distances between *H. elevatus* and *H. pardalinus* butleri were significantly greater than that between *H. elevatus* and *H. pardalinus sergestus.*

Dist. within species	P Δ S	P Δ E	E Δ S	Obsv. r	P-value
0.00001	1	1	1	0.674	0.0001
0.00001	**1**	**1.5**	**1**	**0.782**	**0.0001**
0.00001	1	2	1	0.799	0.0001
0.00001	1	2.5	1	0.789	0.0001
0.00001	1	3	1	0.775	0.0001
0.00001	1	1	1.5	0.521	0.0003
0.00001	1	1	2	0.392	0.0039
0.00001	1	1	2.5	0.302	0.0126
0.00001	1	1	3	0.239	0.0241

Table 5.2 – Results from Mantel tests, comparing Euclidean distance matrix from PCA, to simulated distance matrices with varying distance between *H. pardalinus butleri* and *H. elevatus pseudocupidineus*, as well as *H. pardalinus sergestus* and *H. elevatus pseudocupidineus*.

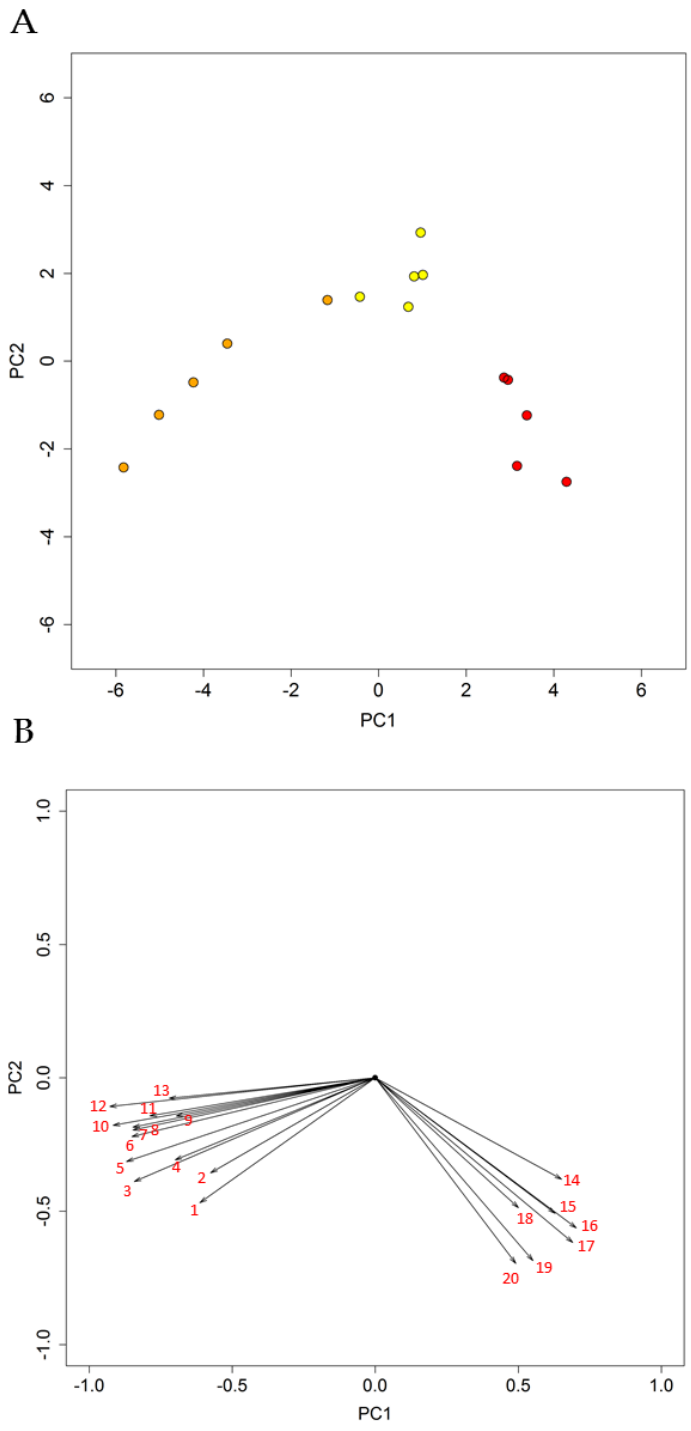


Figure 5.4 – A) Principal component analysis on the 'species pheromone difference dataset'. Colour key; *H. pardalinus butleri* = orange, *H. pardalinus sergestus* = yellow, *H. elevatus pseudocupidineus* = red. B) Variable loadings from Principal component analysis. 1 = heneicosadiene; 2 = homovanillylalcohol; 3 = oleyl acetate; 4 = eicosene; 5 = eicosyl acetate; 6 = (Z)-11-eicosenylpropionate; 7 = phytol; 8 = docosene; 9 = (Z)-9-tricosene; 10 = (Z)-11-eicosenylacetate; 11 = (Z)-9-heneicosene; 12 = hexahydrofarnesyl acetone; 13 = octadecyl acetate; 14 = hexacosanal; 15 = tricosane; 16 = heneicosane; 17 = eicosane; 18 = hexacosane; 19 = 11-methylpentacosane; 20 = unknown heneicosenyl acetate.

Eighteen of the twenty-eight 'candidate pheromones' were found to show significant variation between the species when using Kruskall-Wallis tests, on the data including the York samples, transformed to original relative abundances. When principal component analysis was carried out on this 'York sample abundance dataset' PC1 explained 35.43% of the variance, while PC2 explained 23.16%. Along PC1 *H. pardalinus sergestus* and *H. pardalinus butleri* were not well delimited. However, between *H. elevatus pseudocupidineus* and these two taxa there was clear separation, except for one individual that appears more *H. pardalinus* like. PC2 did not fully delimit any of the taxa, although it did contribute to the separation between *H. pardalinus sergestus* and the others. (Figure 5.5A). Plotting variable loadings this time showed a much wider spread with two compound; hexacosanal and the unknown ketone (Figure 5.5B) clearly contributing more to PC2 than PC1, while most others did not. (Figure 5.5B). Again, other Principal components did not describe variation between the taxa and so were not of interest to this study.

5.4 DISCUSSION

H. elevatus differs strikingly in its colour pattern, from that of its sister species, *H. pardalinus,* as well as most of its other closest relatives in the silvaniform clade (Dasmahapatra *et al.* 2012). Rather than the orange, black and yellow typical of these species, it instead shares the pattern of butterflies in the dennis-rayed mimicry ring. This appears to be due to introgression between *H. elevatus* and its closest comimic *H. melpomene* (Dasmahapatra *et al.* 2012; Wallbank *et al.* 2016). In this chapter I examine two pre-zygotic barriers between *H. elevatus* and *H. pardalinus* in order to see if they support the hypothesis that introgression may have potentially played a role in their speciation in sympatry.

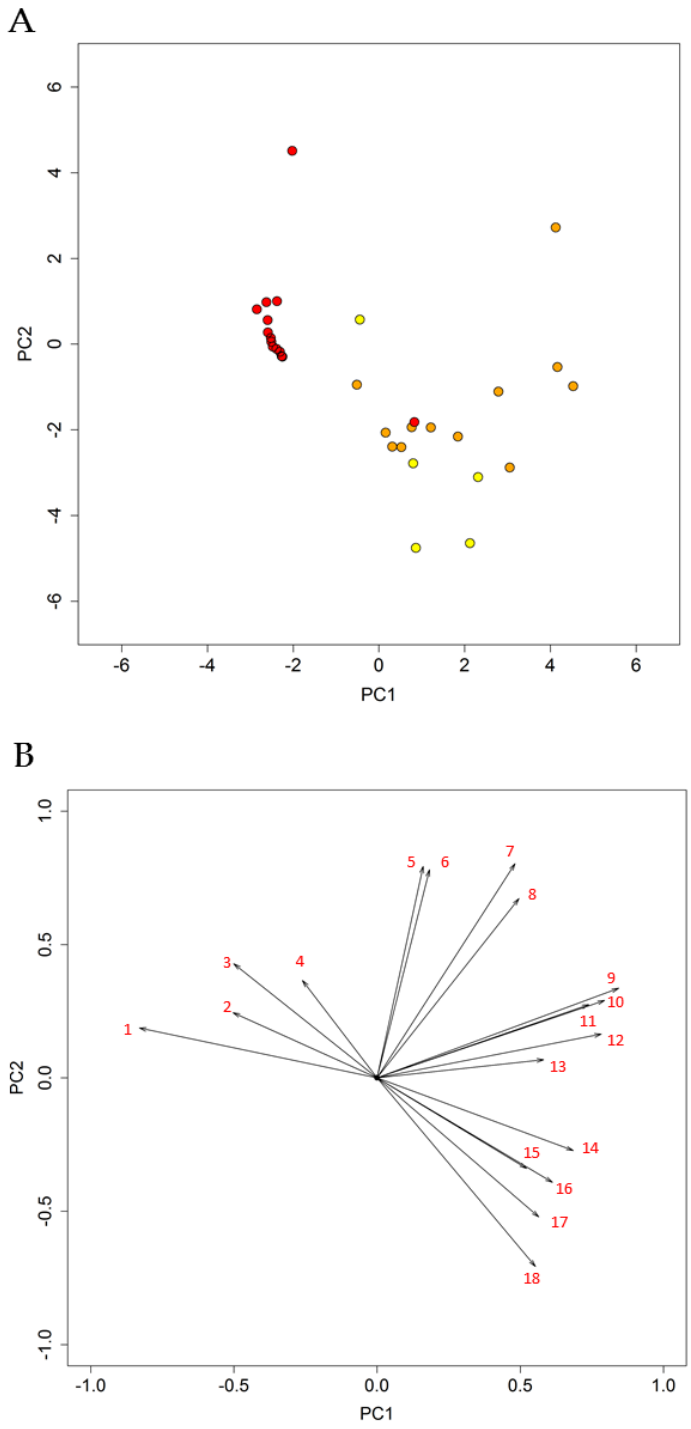


Figure 5.5 – A) Principal component analysis on the 'York sample abundance dataset' which has York samples included. Colour key; *H. pardalinus butleri* = orange, *H. pardalinus sergestus* = yellow, *H. elevatus pseudocupidineus* = red. B) Variable loadings from Principal component analysis. 1 = heneicosane; 2 = tricosane; 3 = eicosane; 4 = 11-methylpentacosane; 5 = hexacosanal; 6 = unknown ketone; 7 = (Z)-11-eicosenal; 8 = (Z)-11-eicosenylacetate; 9 = docosene; 11 = octadecyl acetate; 10 = phytol; 12 = (Z)-9-tricosene; 13 = eicosene; 14 = hexahydrofarnesylacetone; 15 = eicosyl acetate; 16 = (Z)-9-heneicosene; 17 = (Z)-11-eicosenylpropionate; 18 = 11-methyltricosane.

I first investigated the strength of male colour pattern courtship preference for conspecifics, and found it to be significantly different from random, with males showing a preference for conspecifics over heterospecifics. I then investigated putative males sex pheromone differences in both species, using two colour pattern races of *H. pardalinus,* one sympatric with *H. elevatus* and the other allopatric. This revealed a suite of compounds that showed significant differences in quantity between putative male sex pheromone producing regions of male wings and control regions (from corresponding regions from female wings and other regions of male wings). Many of these compounds showed significant differences in quantity between the three taxa, which formed clusters in principal component analsysis, suggesting that all three have differences in there pheromone composition.

5.4.1 **Colour pattern and species discrimination**

The low divergence across most of the genomes of *H. elevatus* and *H. pardalinus* (Kryvokhyzha 2014) can most likely be explained by two rival scenarios. In one of these, the two species speciated in sympatry and strong reproductive barriers led to the 'islands of divergence' that can be so clearly seen. The second includes a phase of allopatry in which the two species diverged, followed by secondary contact, with the genomes of the two species homogenising, but the species remained intact due to strong reproductive barriers controlled by loci that become 'islands of divergence'. Determining which of these two scenarios is more likely to have occurred is difficult. If speciation did occur in sympatry, this would most likely be driven by strong selection on an ecologically important trait that has a secondary role in mate recognition and sexual selection, rather than through sexual selection on traits involved in mate choice and recognition alone (Arnegard & Kondrashov 2004).

The most likely trait therefore to have caused divergence in sympatry is colour pattern, which in other *Heliconius* species is known to be a so called 'magic trait', under both

ecological selection from predators, and also sexual selection due to its role in mate choice (Merrill *et al.* 2011). In *H. cydno,* and two other species that it can hybridise with, *H. melpomene* (Merrill *et al.* 2011) and *H. pachinus* (Kronforst *et al.* 2006b), results from behavioural experiments using hybrid and back cross individuals have shown tight linkage between colour pattern loci and the loci for colour pattern preference. This sympatric speciation would be possible if the introgression of colour pattern genes between *H. elevatus* and *H. melpomene* played an important role in speciation. If this were the case then colour pattern courtship preference for conspecifics is predicted to be strong between *H. elevatus* and *H. pardalinus.* This linkage between colour pattern genes and colour pattern preference genes is important; connecting ecological selection from mimicry that maintains colour pattern races with changes in mate choice that drive speciation. No other pre-zygotic barriers in *Heliconius* are so far known to have this powerful dual effect.

The results presented demonstrate that colour pattern preference is used as a cue during courtship, as was hypothesised, and is therefore at least one part of the suite of traits involved in reproductively isolating the sister species *H. elevatus* and *H. pardalinus.* However, if the introgression of colour pattern genes between *H. elevatus* and *H. melpomene* played an important role in speciation between *H. elevatus* and *H. pardalinus* it colour pattern preference might be predicted to be stronger than was found. In other *Heliconius* sister species for example the strength of colour pattern courtship preference for conspecifics has been found to be stronger; for example it is estimated to be 0.94 for *H. melpomene* males and 0.81 for *H. cydno* males, based on choice experiments (Merrill *et al.* 2011). Overall the strength of reproductive isolation between *H. melpomene* and *H. cydno* is such that in choice experiments with heterospecifics (Jiggins *et al.* 2001; Mavárez *et al.* 2006) there was found to be no mating, which gives an overall barrier, of strength at least 97% (Jiggins 2008). Likewise, mating between *H. pardalinus butleri* and *H. elevatus* in captivity has also been found to be very infrequent, and absent when there are

conspecifics of the opposite sex present (*Rosser pers. comm.*), this suggests that although colour pattern courtship preference is lower in *H. pardalinus butleri* and *H. elevatus* the overall strength of reproductive isolation is similar in the two comparisons, requiring other traits for species discrimination between *H. pardalinus butleri* and *H. elevatus.* This means that some other barrier is likely to be involved in reproductive isolation between *H. elevatus* and *H. pardalinus,* suggesting that while speciation has occurred with gene flow, it has not been driven solely by ecological selection on colour pattern loci and may have occurred in concert with some geographical isolation.

5.4.2 The role of pheromones

GC-MS analysis clearly demonstrated that extracts from the androconial regions of males contained numerous volatile chemical compounds that were not found in extracts from other parts of male wings, or found in regions of the wings in females that are homologous to male androconia. In addition, extracts from the androconia of males of different species showed significantly more differentiation from one another than those from the same species. All together this provides the first real evidence to suggest that pheromones may play an important role in reproductive isolation between *H. elevatus* and *H. pardalinus*. Furthermore, the degree of differentiation in the comparison between sympatric *H. elevatus pseudocupidineus* and *H. pardalinus butleri,* was greater than that between allopatric *H. elevatus pseudocupidineus* and *H. pardalinus sergestus*. These stronger differences in sympatry relative to allopatry, are consistent with the hypothesis that pheromones may have played a role in the initial divergence of these two species in allopatry, as in sympatry only a 'magic trait' under both ecological and sexual selection can likely drive divergence.

Further work, using behavioural experiments to demonstrate that these differences do affect courtship outcome, and perhaps identifying which compounds have the greatest

effect, would further strengthen this argument. Unfortunately due to timing constraints and availability of stocks, these experiments were beyond the scope of this thesis.

Previous work on the lepidopteran pheromones, has found that that fatty-acid synthesis pathway and enzymes are key to their biosynthesis (Liénard *et al.* 2014). Again in this study I found many of the compounds identified likely originate from this pathway, as well as a couple, Homovanillylalcohol and phytol (and syringaldehyde, which was not found to be different between species), that are likely derived from plants. In the analysis on the 'species pheromone difference dataset' (Figure 5.4), which showed the clearest differences between taxa of the two PC analyses, two main clusters of compound were recovered that explained variation along PC1 and PC2.

The first of these contains a number of alkenes as well as their derivatives. These are derived from unsaturated alkanes via desaturase enzymes, which can work at different points along the compound, with the standard being at the first position, as in docosene and eicosene. However, they can also be introduced at other points of the compound. For example, from the presence of (Z)-9-heneicosene in *H. pardalinus* and its lack in *H. elevatus* and the greater abundance of (Z)-9-tricosene in *H. pardalinus*, it appears that *H. pardalinus* uses a Δ9-desaturase that *H. elevatus* does not. This might well be the same desaturate that can work on both C21 and C23 fatty acids. In addition to these compounds, the compounds (Z)-11- eicosenylacetate, found in both *H. pardalinus butleri* and *H. pardalinus sergestus,* and (Z)-11-eicosenylpropionate found only in *H. pardalinus butleri* can also be seen. This suggests that *H. pardalinus* might have an Δ11-desaturase that *H. elevatus* lacks or does not use. In addition, eicosyl acetate is also found in this cluster, this is quite possibly a compound acted upon by the same acetylation enzyme as that of (Z)-11- eicosenylacetate. The second main cluster was dominated by a variety of alkanes. Hexacosanal an aldehyde related to Hexacosane was also in this group along with 11-methylpentacosane.

The clustering found in this study, as well as work by others (Schulz *et al.*, 2008; Mann *et al*, in prep) supports the hypothesis that in both species fatty acids are being synthesised and then used in a variety of downstream enzymatic pathways to create the varied and different bouquets of *Heliconius* butterflies. This allows simple shifts in enzymatic pathways through regulatory changes in numerous genes to quickly build a very different pheromone bouquet. This makes these pathways a very simple way to achieve reproductive isolation, and further lends support to the hypothesis that pheromones play an important role in speciation in *Heliconius*, including between *H. elevatus* and *H. pardalinus*.

5.4.3 The order of barriers

Heliconius courtship proceeds through a set of stereotypical mating behaviours that can finally lead to copulation (Klein & de Araújo 2010; Merrill *et al.* 2015). This begins when a male first encounters a female. If the female is in flight and the male is interested ,the male will pursue the female till she alights or escapes. On the other hand if the female is already alighted or alternatively once the female does alight, the male will begin inspection and possibly begin to hover over the female. During this stage the female rejection response will often be seen, here the female raises her abdomen almost 90 degrees to the angle of her wings, and extrudes her stink-clubs (Eltringham 1925). However, if she does not, and sometimes despite her doing so, the male will alight and attempt copulation by bending his abdomen towards that of the female. Again at this stage the female may well exhibit the rejection response. It's clear that during this progression, colour pattern is first employed by the male in species discrimination, before a female can reject a male based on his pheromone profile during the hovering and male alightment phase.

Reproductive barriers are broadly split into two classes: pre-zygotic and those that are post-zygotic (Coyne & Orr 2004). This is because the order in which barriers to successful reproduction arise during the life cycle is important. However, even between

pre-zygotic barriers some will be involved in courtship first, and those barriers that occur earlier, contribute more to reproductive isolation in absolute terms (Jiggins 2008). Between *H. cydno* and *H. melpomene*, total reproductive isolation has been calculated, along with the strength of habitat choice (Estrada & Jiggins 2002), colour patter preference (Jiggins *et al.* 2001) and the approximated contribution from post-zygotic isolation (Jiggins 2008). The absolute contribution of pheromone to reproductive isolation was only 1.8% due to its late action during courtship, relatively minor compared to the 66% contributed by habitat segregation and the 32% contributed by colour pattern (Jiggins 2008). In the case of *H. elevatus* and *H. pardalinus* the strength of colour pattern preference is considerably less than that found between *H. cydno* and *H. melpomene*, However, due to its earlier role in courtship it likely still plays an important role in mate choice in *H. elevatus* and *H. pardalinus*. In addition, the current strength and importance of isolating barriers does not necessarily reflect their historical importance, or the order in which they evolved (Coyne & Orr 2004).

Colour pattern preference has a prime position during courtship, making it at first seem likely to be the causative agent of speciation between *H. elevatus* and *H. pardalinus.* However, the results here indicate that other traits, potentially involved in reproductive isolation also show differences between these species. Further, not only do colour pattern and pheromones appear to differ between these species, but habitat and host plant preferences also differ (*Rosser pers. comms*). While it seems likely given the strength of colour pattern preference, that the switch in colour pattern in *H. elevatus* may not have driven speciation alone, to further elucidate the causes of the divergence between these two broadly sympatric species further work is necessary. It would be of interest to look at more populations (including allopatric populations of *H. elevatus* in the Guiana's), as well as to better quantify habitat segregation between the two species, as well as to conduct bioassays identifying the actual compounds used as pheromones in these two species.

6. Concluding Remarks

6.1.1 Identification of novel loci

A number of mimicry genes that control major colour pattern elements have previously been identified in *Heliconius*, using a combination of QTL mapping, genome-wide association studies, and by studying gene expression patterns. *Optix* is known to control red-orange pattern elements (Baxter *et al.* 2008b; Reed *et al.* 2011; Martin *et al.* 2014b); *cortex* to control yellow patterning elements (Nadeau *et al.* 2016); and *WntA* to control melanisation around the forewing band, as well as the broken band in *H. erato* (Martin *et al.* 2012; Gallant *et al.* 2014a). These genes have been found to control colour pattern in both *H. erato* and *H. melpomene*. In addition, these loci and therefore likely these genes, have also been shown to be associated with colour pattern in some races of their closer relatives; *H.cydno, H. timareta, H. hecale* and other silvaniforms for *H. melpomene;* and *H. himera* for *H. erato*. Furthermore, two other modifier loci have also been identified, the K locus on chromosome 1 that controls the switch from white to yellow in some species at which the gene *wingless* is found (Kronforst *et al.* 2006b), and the *Ro* locus on chromosome 13 that is involved in forewing band shape in *H. erato* (Papa *et al.* 2013; Nadeau *et al.* 2014).

In chapter 3, I confirmed using QTL mapping analysis that *WntA* controls medial broken shape variation in *H. melpomene*. Furthermore, I also identify *WntA* as the locus controlling the broken band phenotype, as it has been shown to do in *H. erato* (Papa *et al.* 2013). This confirms that convergent evolution has led to the same gene controlling melanisation in the discal part of the forewing in both *H. melpomene* and *H. erato.* In addition to confirming the role of *WntA* in controlling the broken band phenotype across

taxa, I also identified a locus involved in the switch from red to orange pigmentation in *H. melpomene* on Chromosome 13, this completes the loci controlling major colour pattern switches in *H. melpomene*.

In addition to these loci in *H. melpomene* I also investigated the control of the hindwing rays phenotype in *H. demeter* and *H. aoede*. Across a hybrid zone in both species, with the rays present in one race and not the other, I identified regions of elevated divergence around the gene *optix*, known to control this phenotype in both *H. melpomene* and *H. erato*. However, more strikingly I also identified loci across these intraspecific hybrid zones that showed much higher levels of divergence. In the case of *H. demeter* the function of the genes around this locus could not be identified. However, in *H. aoede* the divergence peak was firmly centred on an ommochrome pathway gene, making it an excellent candidate for colour pattern control.

6.1.2 The two-step model and mimicry modifiers

Theoretical models of adaptation suggest that only one or a few loci should account for most of the variation in any given adaptive walk towards a phenotypic optimum, with larger effect mutations being substituted earlier and smaller effect mutations evolving subsequently after (Orr 2005). This two-step model has been hypothesised in mimicry theory for some time, with large effect mutations that cause a mimetic shift thought to evolve first, followed by modifier loci that then refine mimicry (Turner 1977, 1981). In *Heliconius* a handful of large effect genes, *cortex, optix* and *WntA* have now been identified that control switches in colour pattern variation. However, evidence of putative modifier loci that control quantitative variation have also been found in both *H. erato* and *H. melpomene* (Baxter *et al.* 2008a; Papa *et al.* 2013).

In my own QTL analysis I identify a number of putative modifier loci that appear to play roles in medial forewing band variation. While it seems likely that some of these modifiers are unique, a number of these modifier loci mapped to the chromosomes of known major

colour pattern loci. It is possible therefore that some of these major loci may have dual effects in colour patterning, acting as modifier loci for some traits and major switches for others. This dual effect was most clear for the *Ro*/*Or* locus that is known to control medial band shape variation in both *H. erato* (Nadeau *et al.* 2014) and *H. melpomene* (this study). However, as previously mentioned, I identified this locus on chromosome 13 which is known to control the switch from red to orange pigmentation in *H. melpomene*. Two different scenarios can explain this; i) the same gene affects different aspects of mimicry, and ii) two different linked genes each affect different aspects of mimetic colour patterning.

In Nadeau et al. (2014) strong divergence was found at the three known major-effect mimicry loci across both the *H. melpomene* and *H. erato* hybrid zones. However, in both species a number of other putative modifiers were also found. Furthermore, these modifier loci were found to differ between species. It has been hypothesised that this might indicate that while the major mimicry switches are convergent between species, evolution may have led to different modifiers (Kronforst & Papa 2015). In my own analysis of parallel hybrid zones in *H. aoede* and *H. demeter*, I identified elevated divergence in both species at the major pattern switch gene, *optix*. However, in *H. aoede* I also identify much greater divergence at an ommochrome signalling pathway gene *cardinal*, making it an excellent candidate for the genetic control for the loss of rays. It is plausible therefore that in this species both of these two genes contribute to variation in colour pattern across this hybrid zone, with *optix* conserved across species and this novel role for *cardinal* unique to *H. aoede*.

6.1.3 **Ancient pathways, novel functions**

Genes previously identified as controlling colour patterns in *Heliconius* have been found to have conserved roles in other more fundamental developmental pathways. *Cortex* is thought to be involved in cell cycle regulation (Nadeau *et al.* 2016), *optix* in scale

differentiation (Reed *et al.* 2011; Martin *et al.* 2014b), while *wingless* and *WntA* are from the Wnt family of signalling molecules that have been found to be involved in wing development across taxa (Martin & Reed 2010; Lento *et al.* 2013). Given the strategic placement of these patterning genes in developmental pathways involved in wing patterning, where they can control downstream processes involved in pigmentation and scale structure, their potential dual effects are perhaps not surprising (Merrill *et al.* 2015). Furthermore, given the potential dual effects conferred by the *Ro*/*Or* locus in *H. melpomene*, it can be further supposed that the gene at the focal point of this locus may also have a similar effect and placement in butterfly wing patterning pathways.

Ommochrome pigments play a conserved role across insect taxa, controlling pigments that tune insect eyes to natural light conditions (Stavenga 2002). Conserved genes controlling these pigments have been identified through *Drosophila* eye mutants (Haffter *et al.* 1996). In *Heliconius* many of the genes have now been found to be expressed during wing development (Reed & Nagy 2005; Reed *et al.* 2008; Ferguson *et al.* 2011; Hines *et al.* 2012). This is because the orange and red pigments that pattern the wings of *Heliconius* butterflies are also ommochrome pigments (Gilbert 2002; Reed & Nagy 2005; Reed *et al.* 2008). However, previously these genes have not been found to be colour pattern switches (Joron *et al.* 2006a), with instead genes like *optix*, *cortex* and *WntA* controlling colour pattern (Reed *et al.* 2011; Martin *et al.* 2012, 2014b; Nadeau *et al.* 2016). *Cardinal* plays an ancient, conserved role in this fundamental ommochrome pathway, making it a strong candidate for a role in colour pattern control, supported by the high divergence across the *H. aoede* hybrid zone. Furthermore it is associated with patterning in the silkmoth *Bomyx mori*, in which a *cardinal* mutant lacks red pigmentation on the epidermis of final Instar larvae (Osanai-Futahashi *et al.* 2016). My work therefore is concordant with the finding that mimicry genes tend to play conserved roles in developmental pathways that appear to have been co-opted for novel mimicry patterning functions.

6.1.4 **Modulation and enhancer shuffling**

Cis-regulatory enhancer sequences appear to be the main driving force behind much rapid morphological evolution (Wittkopp & Kalay 2012). Individual genes can have multiple enhancers, with each controlling the expression of a gene in a different cell type or at time of development. Through enhancer evolution, genes and developmental pathways can therefore be co-opted, or alternatively assembled de novo into new pathways, for novel functions, while at the same time the function of these genes can be conserved across other developmental networks (Monteiro & Podlaha 2009). This appears to have been the driving force behind mimicry evolution.

Around *cardinal* fixed differences were concentrated not in exons but in introns. While many of these fixed differences are unlikely to be functional but caused by hitchhiking, the lack of divergence in protein coding sequence suggests that coding sequence is conserved across colour pattern forms. Instead it seems likely that changes in *cis*-regulatory modifiers have led to *cardinal's* novel role in the loss of the rays phenotype in *H. aoede*. Furthermore in chapter 4, existing putative regulatory modules were refined around patterning genes, while new ones were also identified, in the species, *H. melpomene*, *H. elevatus* and *H. pardalinus*. These were identified due to their shared derived ancestry across species between races with shared phenotypes, relative to races lacking these shared phenotypes.

The most likely explanation for this shared ancestry across species is through introgression, a signal of which had previously been seen at some loci (Dasmahapatra *et al.* 2012; Pardo-Diaz *et al.* 2012), in this thesis I show for the first time that this signal of introgression is also seen between *H. elevatus* and *H. melpomene* around *WntA*. Furthermore, I identify an even more complex picture of 'enhancer shuffling' (Wallbank *et al.* 2016) between these two species, as well as *H. pardalinus*, with different loci going in different directions. While this had previously been shown for putative *rays* and *dennis*

modules (Wallbank *et al.* 2016), I identify the direction of introgression for putative loci involved in regulating the *optix* forewing band, as well as putative *cortex* and *WntA* enhancers. However, I also identify that in specific cases introgression between species does not appear to have led to colour pattern matching between *H. melpomene* and *H. elevatus*, with convergent evolution instead driving this mimicry.

6.1.5 **Conclusion**

Stochasticity is inbuilt into many aspects of evolution, a sentiment embodied by Stephen J. Gould's (1990) thought experiment of replaying the 'tape of life', in which he postulated that this would result in a different outcome each time. However, it is increasingly apparent that while on a macro scale stochasticity would result in vastly different outcomes, a large proportion of the time when similar traits evolve independently in different species it is through mutations at the same genes (Conte *et al.* 2012). The convergent evolution of mimicry in *Heliconius* has made this system a model for exploring and understanding convergent evolution. However, much of the work on *Heliconius* has focussed on a small number of species and a small number of major effect loci.

The availability of new sequencing technologies opens up the study of the other 30 or so species of *Heliconius* that until now have been largely ignored.. In this thesis I identify novel loci involved in colour patterning and explore the roles of independent evolution, convergent evolution, and introgression, in both the evolution of mimicry and speciation in *Heliconius*. This reveals both interesting cases of convergent genetic evolution, where introgression might have been assumed; as well as cases of independent genetic evolution, where convergent genetic evolution might have been assumed. Broadening the phylogenetic scope of *Heliconius* research in the future will allow us to further understand the repeatability of evolution in *Heliconius* and beyond.

Appendices

Appendix 1 – Sample information from whole genome sequenced individuals from all analyses.

Subspecies	Sample no.	Lat.	Long.	Seq. centre	Platform	Source	Accession no.
H. a. astydamia	MJ09_4015	4° 23' 22" N	52° 12' 36" W	FAS Harvard	HiSeq	-	-
H. a. astydamia	MJ09_4043	4° 32' 42" N	52° 8' 20" W	FAS Harvard	HiSeq	-	-
H. a. astydamia	MJ09_4139	4° 03' 00" N	52° 24' 36" W	FAS Harvard	HiSeq	-	-
H. a. astydamia	MJ09_4197	4° 03' 00" N	52° 24' 36" W	FAS Harvard	HiSeq	-	-
H. a. centurius	2014-47	4° 43' 16" N	56° 48' 35" W	FAS Harvard	HiSeq	-	-
H. a. centurius	2014-97	4° 49' 33" N	57° 24' 02" W	FAS Harvard	HiSeq	-	-
H. a. centurius	CAM021221	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. a. centurius	CAM021231	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. a. cupidineus	JM-09-347	5° 58' 18" S	76° 13' 55" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977656
H. d. beebei	MJ09-4012	4° 23' 22" N	52° 12' 36" W	FAS Harvard	HiSeq	-	-
H. d. beebei	MJ09-4091	4° 19' 27" N	52°49' 12" W	FAS Harvard	HiSeq	-	-
H. d. beebei	MJ09-4115	4° 03' 00" N	52° 24' 36" W	FAS Harvard	HiSeq	-	-
H. d. beebei	MJ09-4164	4° 23' 22" N	52° 12' 36" W	FAS Harvard	HiSeq	-	-
H. d. bouqueti	KD-2014-59	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. d. bouqueti	CAM021201	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. d. bouqueti	CAM021220	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. d. bouqueti	CAM021228	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq		
H. d. demeter	JM-09-323	6° 27' 42" S	76° 17' 30" W	FAS Harvard	HiSeq	Discovar genome	-
H. ele. bari	MJ09-4037	4° 32' 42" N	52° 8'20" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977670
H. ele. bari	MJ09-4056	4° 32' 42" N	52° 8'20" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977671
H. ele. bari	MJ09-4094	4° 32' 42" N	52° 8'20" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977672

H. ele. bari	MJ09-4014	4° 38' 9" N	52° 21' 38" W	FAS Harvard	HiSeq	-	-
H. ele. pseudo.	JM-09-118	5° 54' 37" S	6° 13' 33" W	FAS Harvard	HiSeq	Dasmahapatra et al, 2012	ERS070236
H. ele. pseudo.	JM-09-163	6° 10' 37" S	76° 15' 24" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977673
H. ele. pseudo.	JM-09-270	5° 58' 18" S	76° 13' 55" W	FAS Harvard	HiSeq	Wallbank et al, 2016	*ERS977674*
H. ele. pseudo.	JM-09-302	6° 27'43" S	76° 17' 31" W	FAS Harvard	HiSeq	Dasmahapatra et al, 2012	ERS070238
H. ele. tumatumari	KD-2014-69	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. ele. tumatumari	KD-2014-72	4° 43' 16" N	56°48' 35" W	FAS Harvard	HiSeq	-	-
H. eth. aerotome	JM-09-62	6° 28' 0" S	76° 20' 5" W	FAS Harvard	HiSeq	Wallbank et al, 2016	*ERS977677*
H. h. felix	JM-09-164	6° 10' 37" S	76° 15' 24" W	FAS Harvard	HiSeq	Wallbank et al, 2016	*ERS977681*
H. h. felix	JM-09-273	5° 58' 18" S	76° 13' 54" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235670
H. m. amaryllis	JM-09-216	5° 40' 32" S	77°40' 29" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235653
H. m. amaryllis	JM-11-160	6° 28' 6" S	76° 21' 11" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235652
H. m. amaryllis	JM-11-293	6° 28' 13" S	76° 20' 50" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235654
H. m. meriana	CAM013819	3° 40' 8" N	54° 3' 53" W	GenePool	HiSeq	Wallbank et al, 2016	ERS977703
H. m. meriana	CAM013715	3° 40' 8" N	54° 3' 53" W	GenePool	HiSeq	Wallbank et al, 2016	ERS97774
H. m. rosina	CAM002071	9° 7' 9" N	79°41' 51" W	GenePool	GAII	Dasmahapatra et al, 2012	ERS074426
H. m. rosina	CAM000531	9° 7' 9" N	79°41' 51" W	GenePool	GAII	Martin et al, 2013	ERS235641
H. m. rosina	CAM000533	9° 7' 9" N	79°41' 51" W	GenePool	GAII	Martin et al, 2013	ERS235642
H. m. rosina	CAM000546	9° 7' 9" N	79°41' 51" W	GenePool	GAII	Martin et al, 2013	ERS235643
H. m. thelxiopeia	CAM013566	3° 39' 20" N	54° 2' 21" W	GenePool	HiSeq	Wallbank et al, 2016	ERS977708
H. m.aglaope	JM-09-108	5° 54' 37" S	76° 13' 32" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235655
H. m.aglaope	JM-11-572	5° 56' 44" S	76° 14'47" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235658
H. m.aglaope	JM-11-569	5° 56' 44" S	76° 14'43" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235657
H. m.aglaope	JM-09-112	5° 54' 37" S	76° 13' 32" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235656
H. m.amaryllis	JM-11-48	6° 05' 45" N	76° 58' 38" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235651
H. m.melpomene	CAM013435	4° 54' 54" N	52° 25' 12" W	GenePool	GAII	Martin et al, 2013	ERS235648
H. m.melpomene	CAM009315	4° 57' 47" N	52° 25' 12" W	GenePool	GAII	Martin et al, 2013	ERS235645

H. m.melpomene	CAM009316	4° 57' 47" N	52° 25' 12" W	GenePool	GAII	Martin et al, 2013	ERS235646
H. m.melpomene	CAM009317	4° 57' 47" N	52° 25' 12" W	GenePool	GAII	Martin et al, 2013	ERS235647
H. p. butleri	KD-11-965	6° 17' 53" S	76° 16' 36 " W	FAS Harvard	HiSeq	-	-
H. p. butleri	KD-11-835	6° 18' 27" S	76° 15' 28" W	FAS Harvard	HiSeq	-	-
H. p. sergestus	JM-09-202	6° 28' 40" S	76° 21' 6" W	FAS Harvard	HiSeq	Martin et al, 2013	ERS235668
H. p. sergestus	JM-09-201	6° 28' 40" S	76° 21' 6" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977715
H. p. sergestus	JM-09-209	6° 28' 40" S	76° 21' 6" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977716
H. p. sergestus	JM-09-210	6° 28' 40" S	76° 21' 6" W	FAS Harvard	HiSeq	Wallbank et al, 2016	ERS977717

Appendix 2 – Table showing primer pairs (from chapter 2), annealing temperatures, and approximate regions of reference scaffolds covered.

Name	Forward	Reverse	Anneal. temp.	Reference Region
Dem_0101_PP2	GAAATTCCTCAAGCACCATCTCG	ACACATGTCGTCTCTTAGTACAGTT	64	2727420-2727838
Dem_0101_PP4	TTGTTGCGATAGGATGTTAATGATG	ACGCTATACCTCCATTCCCC	62	2726745-2727648
Dem_0606_PP1	TCTGCGTAATCAACCCTGCTT	TTTTTAGCTTTAGTAGCGTAGCGT	55	1177906-1178820
Dem_0801_PP1	GTACACAGCCACAGGGCG	AGCTGAGGTTGGTGTTTAGGAT	64	2784493-2785153
Dem_0801_PP5	GCTTCGAGGCAAACATGAAACT	GCTGAGGTTGGTGTTTAGGA	65	2784664-2785173
Dem_0901_PP1	GGGTCCGTACCTCAAAAGAATACA	AATGCCTTTATTTTCGCTCGGC	65	7845151-7846140
Dem_1601_PP1	GAAAACGAAATGGACGTGACT	GTGGGTCGATGAGGTGCAA	64	298805-299464
Dem_2101_PP1	CCTGCCATTCACCTGGACAC	ATTGTCCCAAGGTCGCTGTC	62	11942191-11942979
Ner_576_PP1	ACTGTTGACAAGAGCATCGC	ACGTTGGCCAATTTTAAGTATCG	63	100915-101756
Ner_266_PP1	GACGTCCTCGAGGGATCTGG	ACGACATTAACAAAGAGATATCTGAGC	63	176352-176915
Ner_075_PP2	TCAACTCTCAAGCCACATACACA	CACCCTTTTAGGAACGTCGGA	Touchdown	786215-787148

Appendix 3 – *H. aoede* PCR sample information, shows for each primer pair whether sequencing was at least partially successful.

Subspecies	Sample no.	Lat.	Long.	Ner_576_PP1	Ner_266_PP1	Ner_075_PP2
H. a. astydamia	MJ09-4016	4° 23' 22" N	52° 12' 36" W	1	1	1
H. a. astydamia	MJ09-4042	4° 32' 42" N	52° 8' 20" W	1	1	1
H. a. astydamia	MJ09-4044	4° 32' 42" N	52° 8' 20" W	1	1	1
H. a. astydamia	MJ09-4045	4° 32' 42" N	52° 8' 20" W	1	0	1
H. a. astydamia	MJ09-4046	4° 32' 42" N	52° 8' 20" W	0	1	1
H. a. astydamia	MJ09-4061	4° 32' 42" N	52° 8' 20" W	1	1	0
H. a. astydamia	MJ09-4063	4° 32' 42" N	52° 8' 20" W	1	1	0
H. a. astydamia	MJ09-4113	4° 03' 00" N	52° 24' 36" W	1	0	1
H. a. astydamia	MJ09-4114	4° 03' 00" N	52° 24' 36" W	1	1	1
H. a. centurius	CAM021232	4°43' 16" N	56°48' 35" W	1	1	1
H. a. cupidineus	KD-09-296	6° 16' 27" S	76° 10' 23" W	1	1	1
H. a. cupidineus	KD-11-734	6° 17' 53 " S	76° 16' 50" W	1	1	1
H. a. cupidineus	KD-11-497	6° 17' 53 " S	76° 16' 48" W	1	1	1
H. a. cupidineus	KD-05-1286	5° 58' 48" S	76° 13' 85" W	1	1	1
H. a. cupidineus	KD-12-33	12° 34' 05 " S	70° 04' 09" W	0	1	0

Appendix 4 – *H. demeter* PCR sample information, shows for each primer pair whether sequencing was at least partially successful.

Subspecies	Sample no.	Lat.	Long.	0101_PP2	0101_PP4	0606_PP1	0801_PP1	0801_PP5	0901_PP1	1601_PP1	2101_PP1
H. d. beebei	MJ09-4057	4° 32' 42" N	52° 8' 20" W	0	1	0	0	0	0	0	1
H. d. beebei	MJ09-4088	4° 32' 42" N	52° 8' 20" W	0	1	1	0	1	0	1	1
H. d. beebei	MJ09-4089	4° 32' 42" N	52° 8' 20" W	0	1	0	1	0	1	0	1
H. d. beebei	MJ09-4090	4° 32' 42" N	52° 8' 20" W	0	0	1	1	1	1	0	1
H. d. beebei	MJ09-4158	4° 01' 12" N	52° 24' 36" W	1	0	0	0	1	1	0	1
H. d. beebei	MJ09-4162	4° 01' 12" N	52° 24' 36" W	0	0	0	1	1	0	0	1
H. d. beebei	MJ09-4163	4° 01' 12" N	52° 24' 36" W	0	0	0	0	0	0	0	1
H. d. beebei	MJ09-4164	4° 01' 12" N	52° 24' 36" W	0	0	0	1	0	0	1	0
H. d. beebei	MJ09-4165	4° 01' 12" N	52° 24' 36" W	1	0	0	1	1	0	0	1
H. d. beebei	MJ09-4166	4° 01' 12" N	52° 24' 36" W	1	0	0	0	1	1	1	0
H. d. beebei	MJ09-4032	4° 32' 42" N	52° 8' 20" W	0	0	0	1	0	0	1	0
H. d. beebei	MJ09-4033	4° 32' 42" N	52° 8' 20" W	1	1	1	0	0	0	0	1
H. d. bouqueti	KD-2014-65	4° 43' 16" N	56° 48' 35" W	0	0	0	0	0	1	1	1
H. d. demeter	KD-09-321	6° 27' 42" S	76° 17' 30" W	1	0	0	0	1	1	0	1

Appendix 5 – Kruskall Wallis chi-square and p-values from compounds in the 'full compound dataset', looking at significance of variation between male androconial extracts and control extracts. Significantly variable compounds were then included in 'candidate pheromone dataset'. *Although (Z)-11-eicosenylisobutyrate was found to be significantly different between controls and androconial extracts only, it was not present in any androconial extracts and so was not included in in 'candidate pheromone dataset'.

Compound	Chi-sq	P-value
(Z)-11-eicosenol	5.27	0.02
(Z)-11-eicosenylacetate	14.22	0.00
(Z)-11-eicosenylbutanoate	1.20	0.27
(Z)-11-eicosenylisobutyrate	4.72	0.03*
(Z)-11-eicosenylpropionate	6.80	0.01
(Z)-11-eicosenal	12.00	0.00
(Z)-13-docosenyl acetate	0.28	0.60
(Z)-9-heneicosene	8.84	0.00
(Z)-9-tricosene	21.75	0.00
11-methylheptacosane	0.80	0.37
11-methylhexacosane	2.59	0.11
11-methylpentacosane	6.43	0.01
11-methyltricosane	5.51	0.02
19-methyleicosyl acetate	1.20	0.27
1-heneicosene	1.20	0.27
1-octadecanol	3.83	0.05
3-methyl-2-butenyl 3-hydroxystearate	1.20	0.27
3-methyl-3-butenyl 3-hydroxystearate	1.20	0.27
diterpen	3.76	0.05
docosane	0.82	0.37
docosene	12.02	0.00
docosyl acetate	1.20	0.27
eicosanal	0.70	0.41
eicosane	10.16	0.00
eicosene	8.44	0.00
eicosyl acetate	5.27	0.02

ethyl oleate	1.20	0.27
geranylfarnesene	1.20	0.27
heneicosadiene	12.00	0.00
heneicosane	11.18	0.00
heptacosane	3.54	0.06
hexacosanal	10.15	0.00
hexacosane	7.89	0.01
hexadecanal	2.63	0.11
hexadecanol	1.20	0.27
hexadecyl acetate	0.02	0.90
hexahydrofarnesol	1.20	0.27
hexahydrofarnesylacetone	14.22	0.00
homovanillylalcohol	16.00	0.00
nonacosane	2.65	0.10
nonadecane	1.67	0.20
octacosanal	9.40	0.00
octacosane	3.69	0.06
octadecyl acetate	5.42	0.02
oleyl acetate	8.43	0.00
pentacosane	0.04	0.84
phytol	13.89	0.00
squalen	3.42	0.07
syringaaldehyde	8.43	0.00
tetracosane	0.42	0.52
tricosane	14.28	0.00
unknown benzyl-derivative	1.89	0.17
unknown compound	2.48	0.12
unknown heneicosenol	1.21	0.27
unknown heneicosenyl acetate	10.15	0.00
unknown ketone	11.57	0.00
second unknown ketone	4.08	0.04

Appendix 6 – Kruskall Wallis chi-square and p-values from compounds in the 'candidate pheromone dataset', looking at significance of variation between male androconial extracts from the three different species. Significantly variable compounds were then included in 'species pheromone difference dataset'.

Compound	Chi-sq	P-value
(Z)-11-eicosenol	5.139	0.077
(Z)-11-eicosenylacetate	12.963	0.002
(Z)-11-eicosenylpropionate	13.291	0.001
(Z)-11-eicosenal	1.411	0.494
(Z)-9-heneicosene	12.59	0.002
(Z)-9-tricosene	10.693	0.005
11-methylpentacosane	6.02	0.049
11-methyltricosane	0.065	0.968
docosene	11.345	0.003
eicosane	6.27	0.043
eicosene	9.758	0.008
heneicosadiene	7.429	0.024
heneicosane	11.26	0.004
hexacosanal	6.195	0.045
hexacosane	9.986	0.007
hexahydrofarnesylacetone	12.465	0.002
homovanillylalcohol	6.036	0.049
eicosyl acetate	9.912	0.007
octacosanal	3.686	0.158
octadecyl acetate	6.669	0.036
oleyl acetate	7.679	0.022
phytol	10.039	0.007
second unknown ketone	1.734	0.42
syringaaldehyde	0.887	0.642
tricosane	9.707	0.008
unknown heneicosenyl acetate	9.874	0.007
unknown ketone	1.194	0.55

Appendix 7 – Kruskall Wallis chi-square and p-values from compounds in the 'candidate pheromone dataset' when York samples included and abundances converted into relative abundances, and looking at significance of variation between male androconial extracts from the three different species. Significantly variable compounds were then included in 'York sample abundance dataset'.

Compound	Chi-sq	P-value
(Z)-11-eicosenol	2.741	0.25
(Z)-11-eicosenylacetate	28.69	0.00
(Z)-11-eicosenylpropionate	20.62	0.00
(Z)-11-eicosenal	10.28	0.01
(Z)-9-heneicosene	22.73	0.00
(Z)-9-tricosene	15.19	0.00
11-methylpentacosane	13.34	0.00
11-methyltricosane	23.57	0.00
docosene	12.27	0.00
eicosane	18.87	0.00
eicosene	6.20	0.04
heneicosadiene	4.65	0.12
heneicosane	22.758	0.00
hexacosanal	8.31	0.02
hexacosane	5.12	0.08
hexahydrofarnesylacetone	23.67	0.00
homovanillylalcohol	1.38	0.50
eicosyl acetate	6.65	0.04
octacosanal	4.55	0.10
octadecyl acetate	8.45	0.01
oleyl acetate	5.46	0.07
phytol	11.84	0.00
second unknown ketone	9.819	0.01
syringaaldehyde	0.962	0.62
tricosane	23.53	0.00
unknown heneicosenyl acetate	3.10	0.21
unknown ketone	2.83	0.24

Appendix 8 – Script for removal of contigs below or above given values

```
#! /usr/bin/perl -w
use strict;
use warnings;

#removes contigs from fasta file if above or below given sizes

my $length;
my $contig_info;
my $contig;
my $sorted_contigs;

# open contig file
open (CONTIGS, "<$ARGV[0]") or die "could not open contig file.\n";
#Parameter setting
my $min_length = $ARGV[1];
my $max_length = $ARGV[2];

# Output file
$sorted_contigs = "$ARGV[0]_rm_below_$min_length";
open (SORTED_CONTIGS, ">$sorted_contigs");
my $rejected_contigs = 0;
my $retained_contigs = 0;

# sort contigs
while ($contig_info = <CONTIGS>) {
        $contig = <CONTIGS>;
        $length = length($contig);
        if ($length < $min_length or $length > $max_length) {
                $rejected_contigs++;
        }
        else {
                print SORTED_CONTIGS "$contig_info$contig";
            $retained_contigs++;
                }
}

print "Total no. of rejected contigs = $rejected_contigs\n";
print "Total no. of retained contigs = $retained_contigs\n";

close CONTIGS;
close SORTED_CONTIGS;
```

Appendix 9 – Script for removal of contigs below or above given values

```
#! /usr/bin/perl -w
use strict;
use warnings;

# This script takes one ( of the split) contig fasta files as an imput and blasts this against a blastdb made from that file. It also then takes the results as blast output 6 and refines table
# so it contains no overlapping or contained within hits, it also only does each possible pair once.

my $contig_info;
my $contig;
my @contig_info2;
my $contig_size2;
my $blast_reciprocal_temp = "$ARGV[0]_blast_reciprocal_temp";
my $linecounter = 0;

open (INPUTFILE, "<$ARGV[0]") or die "could not open contig file.\n";

#Output file 2
my $reciprical_blast_tab = "$ARGV[0]_reciprical_blast_tab";
open (OUTPUT2, ">$reciprical_blast_tab");

while ($contig_info = <INPUTFILE>){
	$contig = <INPUTFILE>;
	@contig_info2 = split(' ', $contig_info);
	$contig_size2 = $contig_info2[1];

	#Output file 2
	my $prelim_tab_temp = "$ARGV[0]_prelim_tab_temp";
	open (TEMP3, ">$ARGV[0]_prelim_tab_temp");

	#temp contig file
	my $contig_tmp_file = "$ARGV[0]_contig_tmp_file";
	open (TEMP, ">$ARGV[0]_contig_tmp_file");

	print TEMP "$contig_info$contig";

	## open blast res table temp
	  open (TEMP2, ">$ARGV[0]_blast_reciprocal_temp") or  die "could not open blast reciprical temp\n";

	system("/usr/local/src/ncbi-blast-2.2.27+/bin/blastn -task blastn -db 11-569_k50_blastdb.fasta  -query $ARGV[0]_contig_tmp_file -outfmt 6 -max_target_seqs 3 > $ARGV[0]_blast_reciprocal_temp");

	close TEMP;
	close TEMP2;

	#call subroutine
	blasthit_check($blast_reciprocal_temp);
```

```
        close TEMP3;
        system("rm $ARGV[0]_blast_reciprocal_temp");

        #call subroutine
        overlap_check($prelim_tab_temp);

        #remove temp file
        system("rm $ARGV[0]_contig_tmp_file");
        system("rm $ARGV[0]_prelim_tab_temp");

}
close INPUTFILE;
close OUTPUT2;

#checks for overlaps, only outputs non-overlapping hits

sub overlap_check {

my $prelim_tab_temp = $_[0];
open (INPUT3, "<$ARGV[0]_prelim_tab_temp") or  die "could not open
blast_results_file_temp for read\n";

my @blast_info3;
my $blast_table3;
my $blast_table2;
my @blast_info2;
my $counter = 1;

##print to final blast table
my @startq = ();
my @endq = ();
my @starth = ();
my @endh = ();

my $linecounter4 = 0;
my $print_or_not = 0;

#change split to tab
#read in first line of the blast table
while ($blast_table3 = <INPUT3>) {
        @blast_info3 = split(' ', $blast_table3);
        if (($blast_info3[6] < $blast_info3[7]) && ($blast_info3[8] < $blast_info3[9])){
          push @startq, $blast_info3[6];
        push @endq, $blast_info3[7];
          push @starth, $blast_info3[8];
        push @endh, $blast_info3[9];
          }
          elsif (($blast_info3[6] > $blast_info3[7]) && ($blast_info3[8] < $blast_info3[9])){
          push @startq, $blast_info3[7];
        push @endq, $blast_info3[6];
        push @starth, $blast_info3[8];
        push @endh, $blast_info3[9];
          }
          elsif (($blast_info3[6] < $blast_info3[7]) && ($blast_info3[8] > $blast_info3[9])){
```

```
        push @startq, $blast_info3[6];
      push @endq, $blast_info3[7];
      push @starth, $blast_info3[9];
      push @endh, $blast_info3[8];
        }
        else{
        push @startq, $blast_info3[7];
      push @endq, $blast_info3[6];
      push @starth, $blast_info3[9];
      push @endh, $blast_info3[8];
        }
      print OUTPUT2
"$blast_info3[0]\t$blast_info3[1]\t$blast_info3[2]\t$blast_info3[3]\t$blast_info3[4]\t$blast
_info3[5]\t$blast_info3[6]\t$blast_info3[7]\t$blast_info3[8]\t$blast_info3[9]\t$blast_info3
[10]\t$blast_info3[11]\n";
        #start loop to read in all other lines add starts and ends to arrays
      while ($blast_table2 = <INPUT3>) {
            @blast_info2 = split(' ', $blast_table2);
                if (($blast_info2[6] < $blast_info2[7]) && ($blast_info2[8] <
$blast_info2[9])){
              push @startq, $blast_info2[6];
              push @endq, $blast_info2[7];
              push @starth, $blast_info2[8];
        push @endh, $blast_info2[9];
        }
        elsif (($blast_info2[6] > $blast_info2[7]) && ($blast_info2[8] < $blast_info2[9])){
        push @startq, $blast_info2[7];
        push @endq, $blast_info2[6];
        push @starth, $blast_info2[8];
        push @endh, $blast_info2[9];
        }
        elsif (($blast_info2[6] < $blast_info2[7]) && ($blast_info2[8] > $blast_info2[9])){
        push @startq, $blast_info2[6];
        push @endq, $blast_info2[7];
        push @starth, $blast_info2[9];
                push @endh, $blast_info2[8];
        }
                else{
        push @startq, $blast_info2[7];
        push @endq, $blast_info2[6];
        push @starth, $blast_info2[9];
        push @endh, $blast_info2[8];
        }
                $linecounter4 ++;
                $counter = 1; #this counter is used in order  to compare to all previous
hits
                $print_or_not = 0; #value added to if breaks conditions and means it is
not printed to new table
                while ($counter <= $linecounter4) {
                        if ($startq[$linecounter4] >= $startq[($linecounter4 - $counter)]
&& $endq[$linecounter4] <= $endq[($linecounter4 - $counter)] or
$startq[$linecounter4] < $endq[($linecounter4 - $counter)] && $endq[$linecounter4] >
$endq[($linecounter4 - $counter)] or $endq[$linecounter4] > $startq[($linecounter4 -
$counter)] && $startq[$linecounter4] < $startq[($linecounter4 - $counter)] or
```

```
$starth[$linecounter4] >= $starth[($linecounter4 - $counter)] && $endh[$linecounter4]
<= $endh[($linecounter4 - $counter)] or $starth[$linecounter4] < $endh[($linecounter4
- $counter)] && $endh[$linecounter4] > $endh[($linecounter4 - $counter)] or
$endh[$linecounter4] > $starth[($linecounter4 - $counter)] && $starth[$linecounter4] <
$starth[($linecounter4 - $counter)]) {
                        $print_or_not ++;
                        pop @starth; #remove last value from array if not printing as other
contig may come that doesn't overlap printed but does overlap unprinted
                pop @endh;
                        pop @startq; #remove last value from array if not printing as other
contig may come that doesn't overlap printed but does overlap unprinted
                        pop @endq; # same as above
                        $linecounter4 = $linecounter4 -1; #same as above
                        last;
                        }
                        else {
                        $counter ++; #if doesn't break condition add one to counter so as
to compare against the next hit
                        }}
                        #print if it hasn't broken conditions and so = 0
                unless ($print_or_not != 0) {
                print OUTPUT2
"$blast_info2[0]\t$blast_info2[1]\t$blast_info2[2]\t$blast_info2[3]\t$blast_info2[4]\t$blast
_info2[5]\t$blast_info2[6]\t$blast_info2[7]\t$blast_info2[8]\t$blast_info2[9]\t$blast_info2
[10]\t$blast_info2[11]\n";
}}}
close INPUT3;

}

##checks whether this pair have already hit

sub blasthit_check {

my $blast_reciprocal_temp = $_[0];
open (INPUT2, "<$ARGV[0]_blast_reciprocal_temp") or  die "could not open
blast_reciprocal_temp for read\n";

my $linecounter = 0;
my $linecounter2 = 0;
my @hit_contig = ();
my $second_hit_contig;
my $blast_table;
my @blast_info;

#change split to tab
#read in first line of the blast table
while ($blast_table = <INPUT2>) {
        $linecounter++;
      @blast_info = split('\t', $blast_table);
      push @hit_contig, $blast_info[1];
        if (($blast_info[0] != $blast_info[1]) && ($linecounter2 == 0)){
                $second_hit_contig =  $blast_info[1];
                $linecounter2++;
```

```
                print TEMP3 "$blast_table";
        }
        elsif (($linecounter2 != 0) && ($blast_info[1] == $second_hit_contig)){
                $linecounter2++;
                print TEMP3 "$blast_table";
        }
        else{
}}
close INPUT2;
}
```

Appendix 10 – Script to find redundant contigs

```
#! /usr/bin/perl -w
use strict;
use warnings;

### finds redundant contigs given a reciprocal blast table from reciprical_blast.pl, outputs
as a list of redundant contigs  ## that have similarities above set threshold


########### make new blast file buy adding contig size to each blast line ############

# open the blast table file
open (INPUTFILE1, "<$ARGV[0]") or die "could not open input blast table file.\n";

# open the contig file
open (INPUTFILE2, "<$ARGV[1]") or die "could not open input contig fasta file.\n";

my $total_num_contigs = 0; my $line1_count = 0; my $line1; my @blast_info;
my @contig_num; my @line2; my $ignore_line; my @contig_line; my $i;

my $output_file1 = "$ARGV[0].blast_with_size$ARGV[2]$ARGV[3]";
open (OUTPUT1, ">$output_file1");

my $output_file2 = "$ARGV[0].bad_contigs$ARGV[2]$ARGV[3]";
open (OUTPUT2, ">$output_file2");

#Percent ID and proportion thresholds
my $ID_threshold = $ARGV[2]; #in form of number
my $prop_threshold = $ARGV[3]; #in form of number

my $prop_threshold2 = ($prop_threshold / 100);

for ($i = 0; $i < 1; ++$i){
       $contig_num[$i] = 0;
}

for ($i = 0; $i < 1; ++$i){
       $line2[$i] = 0;
```

```
}

while ($line1 = <INPUTFILE1>){
	@blast_info =  split(' ', $line1);
	  $line1_count ++;
	$contig_num[$line1_count] = $blast_info[0];
	  if ($contig_num[$line1_count] == $contig_num[$line1_count -1]){
		print OUTPUT1
"$contig_line[1]\t$blast_info[0]\t$blast_info[1]\t$blast_info[2]\t$blast_info[3]\t$blast_inf
o[4]\t$blast_info[5]\t$blast_info[6]\t$blast_info[7]\t$blast_info[8]\t$blast_info[9]\t$blast_
info[10]\t$blast_info[11]\n";
		}
		else {
			while ($line2[$line1_count] = <INPUTFILE2>){
			$ignore_line = <INPUTFILE2>;
			@contig_line = split(' ', $line2[$line1_count]);
				if (">$blast_info[0]" eq "$contig_line[0]"){
					$total_num_contigs++;
				print OUTPUT1
"$contig_line[1]\t$blast_info[0]\t$blast_info[1]\t$blast_info[2]\t$blast_info[3]\t$blast_inf
o[4]\t$blast_info[5]\t$blast_info[6]\t$blast_info[7]\t$blast_info[8]\t$blast_info[9]\t$blast_
info[10]\t$blast_info[11]\n";
				last;
}}}}

close INPUTFILE1;
close INPUTFILE2;
close OUTPUT1;

print "num. of contigs = $total_num_contigs\n";

##########################

# open the new blast file
open (INPUTFILE3, "<$ARGV[0].blast_with_size$ARGV[2]$ARGV[3]") or die "could not
open the new blast table file.\n";

my $contig_count = 0; my $launch = 0; my @contig_num2; my @line4; my
@total_ID_score_for_hit; my @sum_alignment_size; my @contig_size;
my @mean_ID_score_for_hit; my $line3; my @blast_info2; my $line3_count; my
$num_hits; my @proportion_of_hit_aligned;

#arrays for keepng scores in number of 0s same as total number of contigs counted
previously

for ($i = 0; $i < ($total_num_contigs + 1); ++$i){
	$contig_num2[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
	 $line4[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
```

```
	$total_ID_score_for_hit[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
	$sum_alignment_size[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
	$contig_size[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
	$mean_ID_score_for_hit[$i] = 0;
}

for ($i = 0; $i < $total_num_contigs; ++$i){
	$proportion_of_hit_aligned[$i] = 0;
}

my @blast_info2_array;

for ($i = 0; $i < $total_num_contigs; ++$i){
	$blast_info2_array[$i] = 0;
}


############### CALCULATE THE PRLIM SOCRES FOR EACH CONTIG

while ($line3 = <INPUTFILE3>){ #read in file
	@blast_info2 =  split(' ', $line3);
	$line3_count ++;
	  $blast_info2_array[$line3_count] = $blast_info2[2];
	$contig_num2[$line3_count] = $blast_info2[1];
	  $launch++; # this is needed to sort out what to do with first line
	if (($contig_count == 0) && ($contig_num2[$line3_count] !=
$contig_num2[$line3_count -1]) && ($launch == 1)) { #this loop takes first line and adds
to scores in first array elements

			$contig_size[$contig_count] = $blast_info2[0];
			$sum_alignment_size[$contig_count] =
$sum_alignment_size[$contig_count] + $blast_info2[4];
			$total_ID_score_for_hit[$contig_count] =
$total_ID_score_for_hit[($contig_count)] + ($blast_info2[4] * $blast_info2[3]);
			$num_hits = 1;
				$contig_count++;

	  }
	elsif (($contig_count == 0) && ($contig_num2[$line3_count] ==
$contig_num2[$line3_count -1])) { ## this is only used if first contig has more than one
hit

				$contig_size[($contig_count -1)] = $blast_info2[0];
			$sum_alignment_size[($contig_count -1)] =
$sum_alignment_size[($contig_count -1)] + $blast_info2[4];
```

```
                $total_ID_score_for_hit[($contig_count -1)] =
$total_ID_score_for_hit[($contig_count -1)] + ($blast_info2[4] * $blast_info2[3]);
                $num_hits++; ### if contig has multiple hits this counter counts them so
as to be bale to get mean scores for each contig
        }
      else {
            ### calculate scores from previously collected contig
            if (($contig_count >= 1) &&( $contig_num2[($line3_count)] !=
$contig_num2[$line3_count -1])) {
                $mean_ID_score_for_hit[($contig_count -1)] =
($total_ID_score_for_hit[($contig_count -1)] / $sum_alignment_size[($contig_count -
1)]);
            $proportion_of_hit_aligned[($contig_count -1)] =
$sum_alignment_size[($contig_count -1)] / $contig_size[($contig_count -1)];
                if (($mean_ID_score_for_hit[($contig_count -1)] > $ID_threshold)
&& ($proportion_of_hit_aligned[($contig_count -1)] > $prop_threshold2)){

                    print OUTPUT2 "$contig_num2[($line3_count -
1)]\t$blast_info2_array[($line3_count -1)]\t$contig_size[($contig_count -
1)]\t$mean_ID_score_for_hit[($contig_count -
1)]\t$proportion_of_hit_aligned[($contig_count -1)]\n";
                }
            }
            if ($contig_num2[$line3_count] != $contig_num2[$line3_count -1]){

                $contig_size[$contig_count] = $blast_info2[0];
              $sum_alignment_size[$contig_count] =  $blast_info2[4];
              $total_ID_score_for_hit[$contig_count] =
$total_ID_score_for_hit[($contig_count)] + ($blast_info2[4] * $blast_info2[3]);
              $num_hits = 1;
                $contig_count++;
            }
        else {
              $contig_size[($contig_count -1)] = $blast_info2[0];
              $sum_alignment_size[($contig_count -1)] =
$sum_alignment_size[($contig_count -1)] + $blast_info2[4];
              $total_ID_score_for_hit[($contig_count -1)] =
$total_ID_score_for_hit[($contig_count -1)] + ($blast_info2[3] * $blast_info2[4]);
              $num_hits++;
            }
      }
}

### calculate mean ID score and proportion aligned for last contig

#print "hits = $num_hits\n";
$mean_ID_score_for_hit[($contig_count -1)] =
($total_ID_score_for_hit[($contig_count -1)] / $sum_alignment_size[($contig_count -
1)]);

#print "moo\n";
$proportion_of_hit_aligned[($contig_count -1)] = $sum_alignment_size[($contig_count -
1)] / $contig_size[($contig_count -1)];
```

```
#print "$mean_ID_score_for_hit[($contig_count -1)]\n";
if (($mean_ID_score_for_hit[($contig_count -1)] > $ID_threshold) &&
($proportion_of_hit_aligned[($contig_count -1)] > $prop_threshold2)){
	print OUTPUT2 "$contig_num2[($line3_count -
1)]\t$blast_info2_array[($line3_count -1)]\t$contig_size[($contig_count -
1)]\t$mean_ID_score_for_hit[($contig_count -
1)]\t$proportion_of_hit_aligned[($contig_count -1)]\n";
}

close INPUTFILE1;
close INPUTFILE2;
close INPUTFILE3;
close OUTPUT2;
close OUTPUT1;
```

Appendix 11 – Script to make list of smaller redundant contigs

```
#! /usr/bin/perl -w
use strict;
use warnings;

#removes the larger of any pair of redundant contigs and outputs to a list
'only_unique_bad' so that they can then  be removed with filter_redundants.pl

my $line; my $line2; my @info; my $contig_num; my $contig_element0; my $line3; my
@hit_info; my $query_name; my $hit_name; my $query_size; my $key; my $key2;
my $removed_already = 0; my %removed_hash = (); my $hit_size; my $value;
my %size_hash = ();

# open bad contig file
open (INPUTFILE, "<$ARGV[0]") or die "could not open bad contig file.\n";

# open bad contig file
open (INPUTFILE2, "<$ARGV[1]") or die "could not open full contig file.\n";

my $output_file1 = "$ARGV[0]_only_unique_bad";
open (OUTPUT, ">$output_file1");

while ($line = <INPUTFILE2>) {
	$line2 = <INPUTFILE2>;
	@info = split(' ', $line);
	$hit_size = $info[1];
	$contig_element0 = $info[0];
	$contig_num = substr($contig_element0,1);
	$size_hash{$contig_num} = $hit_size;
}
close INPUTFILE2;
open (INPUTFILE2, "<$ARGV[1]") or die "could not open full contig file in loop.\n";

while ($line3 = <INPUTFILE>) {
```

```
        $removed_already = 0;
      @hit_info = split(' ', $line3);
      $query_name = $hit_info[0];
      $hit_name = $hit_info[1];
      $query_size = $hit_info[2];
        for $key ($hit_name) {
              $value = $size_hash{$key};
              if ($value < $query_size){
                    for $key ($hit_name) {
                          $removed_already = 1 if exists $removed_hash{$key};
                    if ($removed_already == 0){
                          print OUTPUT "$hit_name\t$value\n";
                          $removed_hash{$hit_name} = $value;
                          }
                    }
              }
              elsif ($value == $query_size){
                    for $key2 ($query_name) {
                          $removed_already = 1 if exists $removed_hash{$key};
                    if ($removed_already == 0){
                          print OUTPUT "$query_name\t$query_size\n";
                          $removed_hash{$query_name} = $query_size;
                    }
                    }
            }
              else {
                    for $key2 ($query_name) {
                          $removed_already = 1 if exists $removed_hash{$key2};
                    if ($removed_already == 0){
                          print OUTPUT "$query_name\t$query_size\n";
                          $removed_hash{$query_name} = $query_size;

}}}}}
```

Appendix 12 – Script takes list of redundant contigs to be removed, and removes them from assembly

```
#! /usr/bin/perl -w
use strict;
use warnings;

# Takes unique_bad file and removes all of these from original assembly fasta to make a
new filtered_redundant_contig # file, again put in thresholds

my $line; my $line2; my @info; my $contig_element0; my $contig_num; my
%contig_hash;
my $line3; my %redundants_hash; my %info_hash;
my $key;
my $value;
```

```
my $key2;
my $value2;

# open the contig file
open (INPUTFILE1, "<$ARGV[0]") or die "could not open input contig fasta file.\n";

# open the redundant contig list
open (INPUTFILE2, "<$ARGV[1]") or die "could not open redundant contig list
only_unique_bad.\n";

my $ID_threshold = $ARGV[2]; #in form of number
my $prop_threshold = $ARGV[3]; #in form of number

#read in contig file to hash, read in list to hash, then do if exists print

my $output_file1 = "$ARGV[0]_filtered_contigs$ARGV[2]$ARGV[3]";
open (OUTPUT, ">$output_file1");

#put redundants into hash
while ($line3 = <INPUTFILE2>) {
	while ($line3 = <INPUTFILE2>) {
		@info = split(' ', $line3);
		$contig_num = $info[0];
		$redundants_hash{$contig_num} = $contig_num;
}}


###put contig file into hash, query as key and then hit as value in one size as value in
other

while ($line = <INPUTFILE1>) {
	$line2 = <INPUTFILE1>;
	chomp $line2;
	chomp $line;
	@info = split(' ', $line);
	$contig_element0 = $info[0];
	$contig_num = substr($contig_element0,1);
	$contig_hash{$contig_num} = $line2;
	$info_hash{$contig_num} = $line;
}

while ( my ($key, $value) = each(%contig_hash) ) {
}

while ( my ($key, $value) = each(%info_hash) ) {
}

close INPUTFILE1;

open (INPUTFILE1, "<$ARGV[0]") or die "could not open input contig fasta file.\n";

foreach $key (keys %contig_hash) {
	(($key, $value) = each(%contig_hash));
	(($key2, $value2) = each(%info_hash));
```

```
        print OUTPUT "$value2\n$value\n" unless exists $redundants_hash{$key};
}
```

Appendix 13 – Map *denovo* contigs with fixed positions to reference genome with Blastn

```
#! /usr/bin/perl -w
use strict;
use warnings;


# blasts contigs with fixed differences and works out the location of each fixed difference
on that scaffold. This can then be mapped to genomic positions.
# input is contig fasta file and contig info file from fixed_diff_per_contig.pl

my $contig_line; my $pos_line; my $pos_count_line; my $ignore; my @contig_num_line;
my $contig_num; my @SNP_positions; my @pos_count; my $pos_num; my $loop_num;
my $lines; my @blast_lines; my $query_start; my $query_end; my $target_start; my
$target_end; my $SNP_ref_position;
my $num_lines; my $loop_counter; my $unmapped_SNP = 0;

## open contig info file
open (INPUTFILE, "<$ARGV[0]") or die "could not open contig info file.\n";

## open fixed diff file
open (INPUTFILE2, "<$ARGV[1]") or die "could not open fixed diff file.\n";

## open contig file of reference
my $ref = "$ARGV[2]";
#open (INPUTFILE1, "<$ARGV[1]") or die "could not open contig file of reference.\n";

## open output main
my $blast_map = "$ARGV[0]_blast_geno_ass_map";
open (OUTPUT1, ">$ARGV[0]_blast_geno_ass_map");

my $blast_map_with_geno_calls = "$ARGV[0].blast_map_with_geno_calls";
open (OUTPUT2, ">$blast_map_with_geno_calls");


#read in contig info file made from freq_table.pl
while ($contig_line = <INPUTFILE>) {
        $pos_line = <INPUTFILE>;
        $pos_count_line = <INPUTFILE>;
        $ignore = <INPUTFILE>;
        @contig_num_line = split(' ',$contig_line);
        $contig_num = $contig_num_line[2];
        @SNP_positions = split('\t',$pos_line);
        @pos_count = split(' ',$pos_count_line);
        $pos_num = ($pos_count[2]);
        $loop_num = 1;
```

```
        #temp contig file
      my $contig_tmp_file = "$ARGV[0]_contig_tmp_file";
      open (TEMP, ">$ARGV[0]_contig_tmp_file");

        my $blast_table_tmp_file = "$ARGV[0]_blast_table_tmp_file";
      open (TEMP2, ">$ARGV[0]_blast_table_tmp_file");

        #grep contig into temporary file
        system ("grep -A 1 '>$contig_num' $ref > $contig_tmp_file");
        #blast this temporary file
        system("/usr/local/src/ncbi-blast-2.2.27+/bin/blastn -task blastn -db Hmel1-
1_primaryScaffolds_mtDNA.fasta  -query $contig_tmp_file -outfmt 6 -max_target_seqs 2
> $blast_table_tmp_file");

        close TEMP2;

        open (TEMP2, "<$ARGV[0]_blast_table_tmp_file") or die "could not open
blast_table_tmp_file for readin";

        $num_lines = 0;
        $loop_counter = 0;

        while ($lines = <TEMP2>) {
                $num_lines++;
        }
        close TEMP2;

      open (TEMP2, "<$ARGV[0]_blast_table_tmp_file") or die "could not open
blast_table_tmp_file for readin";

        while (($loop_num <= $pos_num) && ($loop_counter < $num_lines)) {
                $lines = <TEMP2>;
                $loop_counter++;
                        @blast_lines = split('\t',$lines);
                        if ($blast_lines[0] == $contig_num){
                                $query_start = $blast_lines[6];
                                $query_end = $blast_lines[7];
                                $target_start = $blast_lines[8];
                                $target_end = $blast_lines[9];
                                if (($SNP_positions[$loop_num] >= $query_start) &&
($SNP_positions[$loop_num] <= $query_end) && ($num_lines  >  $loop_counter)){
                                        if ($target_start < $target_end){
                                                chomp $SNP_positions[$loop_num];
                                                print OUTPUT1
"$contig_num\t$SNP_positions[$loop_num]\t$blast_lines[1]\t";
                                                $SNP_ref_position =
(($SNP_positions[$loop_num] + $target_start) - $query_start);
                                                print OUTPUT1 "$SNP_ref_position\n";
                                        }
                                        else {
                                                chomp $SNP_positions[$loop_num];
                                                print OUTPUT1
"$contig_num\t$SNP_positions[$loop_num]\t$blast_lines[1]\t";
```

```
                                        $SNP_ref_position = (($target_start -
$SNP_positions[$loop_num]) + $query_start);
                                        print OUTPUT1 "$SNP_ref_position\n";
                                            }
                                            $loop_counter = 0;
                                            $loop_num++;
                                            close TEMP2;
                                            open (TEMP2, "<$ARGV[0]_blast_table_tmp_file")
or die "could not open blast_table_tmp_file while in loop";
                                        }
                                        elsif ($num_lines == $loop_counter){
                                        $unmapped_SNP++;
                                        $loop_counter = 0;
                                        $loop_num++;
                                        close TEMP2;
                                open (TEMP2, "<$ARGV[0]_blast_table_tmp_file") or die "could
not open blast_table_tmp_file while in loop";
                                        }
                                }
                        }
                }
        }
}

close TEMP2;
close TEMP;
system ("rm $ARGV[0]_blast_table_tmp_file");
system ("rm $ARGV[0]_contig_tmp_file");

#call subroutine
make_blast_map_with_genos($blast_map);

system ("rm $ARGV[0]_blast_geno_ass_map");

##############################################################################

sub make_blast_map_with_genos {

my $blast_map = $_[0];
open (INPUTFILE3, "<$ARGV[0]_blast_geno_ass_map") or  die "could not open
blast_geno_ass_map\n";

my $line_fixdiff; my $line_map; my @map_info; my @fixdiff_info; my $line_take = 0;

##make a new file that has blast map coords and genotype calls on it

while ($line_fixdiff = <INPUTFILE2>){
        while ($line_fixdiff = <INPUTFILE2>) {
                if ($line_take == 0) {
                        $line_map = <INPUTFILE3>;
                        @map_info = split('\t', $line_map);
                        @fixdiff_info = split('\t', $line_fixdiff);
                        if (($map_info[0] == $fixdiff_info[0]) && ($map_info[1] ==
$fixdiff_info[1])) {
                                chomp $map_info[3];
```

```
                        chomp $fixdiff_info[9];
                        print OUTPUT2
"$map_info[2]\t$map_info[3]\t$fixdiff_info[2]\t$fixdiff_info[3]\t$fixdiff_info[4]\t$fixdiff_inf
o[5]\t$fixdiff_info[6]\t$fixdiff_info[7]\t$fixdiff_info[8]\t$fixdiff_info[9]\n";
                        $line_take = 0;
                }
                else {
                        $line_take = 1;
                        #do_nothing
                }
        }
        else {
                @map_info = split('\t', $line_map);
                @fixdiff_info = split('\t', $line_fixdiff);
                if (($map_info[0] == $fixdiff_info[0]) && ($map_info[1] ==
$fixdiff_info[1])) {
                        chomp $map_info[3];
                        chomp $fixdiff_info[9];
                        print OUTPUT2
"$map_info[2]\t$map_info[3]\t$fixdiff_info[2]\t$fixdiff_info[3]\t$fixdiff_info[4]\t$fixdiff_inf
o[5]\t$fixdiff_info[6]]\t$fixdiff_info[7]\t$fixdiff_info[8]\t$fixdiff_info[9]\n";
                        $line_take = 0;
                }
                else {
                        $line_take = 1;
                        #do_nothing
                }
        }
}}}
```

Appendix 14 – Script to run Blasts for redundancy estimation

```
#! /usr/bin/perl -w
use strict;
use warnings;

# this blasts a contig database against reference to give all non overlapping (in hit or
query) results for # each contig, against reference genome. Can then be used to work out
redundancy in genome

my $contig_info;
my $contig;
my @contig_info2;
my $contig_size2;
my $blast_redundancy_temp = "$ARGV[0]_blast_redundancy_temp";
my $linecounter = 0;

open (INPUTFILE, "<$ARGV[0]") or die "could not open contig file.\n";
```

```
#Output file 2
my $redundancy_blast_tab = "$ARGV[0]_redundancy_blast_tab";
open (OUTPUT2, ">$redundancy_blast_tab");

while ($contig_info = <INPUTFILE>){
	$contig = <INPUTFILE>;
	@contig_info2 = split(' ', $contig_info);
	$contig_size2 = $contig_info2[1];

	  #Output file 2
	my $prelim_tab_temp = "$ARGV[0]_prelim_tab_temp";
	open (TEMP3, ">$ARGV[0]_prelim_tab_temp");

	#temp contig file
	my $contig_tmp_file = "$ARGV[0]_contig_tmp_file";
	open (TEMP, ">$ARGV[0]_contig_tmp_file");

	print TEMP "$contig_info$contig";

	## open blast res table temp
	  open (TEMP2, ">$ARGV[0]_blast_redundancy_temp") or  die "could not open
blast redundancy temp\n";

	system("/usr/local/src/ncbi-blast-2.2.27+/bin/blastn -task blastn -db Hmel1-
1_primaryScaffolds_mtDNA.fasta  -query $ARGV[0]_contig_tmp_file -outfmt 6 -
max_target_seqs 2 > $ARGV[0]_blast_redundancy_temp");

	close TEMP;
	close TEMP2;

	  #call subroutine to remove last line from raw blast output file
	remove_last_line_of_tab($blast_redundancy_temp);

	close TEMP3;
	system("rm $ARGV[0]_blast_redundancy_temp");

	#call subroutine to print to final table only non-overlapping (in hit or query coords)
results for each query
	overlap_check($prelim_tab_temp);

	#remove temp file
	system("rm $ARGV[0]_contig_tmp_file");
	system("rm $ARGV[0]_prelim_tab_temp");

}

close INPUTFILE;
close OUTPUT2;

################################################################################
sub overlap_check {

my $prelim_tab_temp = $_[0];
```

```
open (INPUT3, "<$ARGV[0]_prelim_tab_temp") or  die "could not open
blast_results_file_temp for read\n";

my @blast_info3;
my $blast_table3;
my $blast_table2;
my @blast_info2;
my $counter = 1;

##print to final blast table
my @startq = ();
my @endq = ();
my @starth = ();
my @endh = ();

my $linecounter4 = 0;
my $print_or_not = 0;

#change split to tab
#read in first line of the blast table
while ($blast_table3 = <INPUT3>) {
	@blast_info3 = split('\t', $blast_table3);
	if (($blast_info3[6] < $blast_info3[7]) && ($blast_info3[8] < $blast_info3[9])){
	  push @startq, $blast_info3[6];
	push @endq, $blast_info3[7];
	  push @starth, $blast_info3[8];
	push @endh, $blast_info3[9];
	  }
	  elsif (($blast_info3[6] > $blast_info3[7]) && ($blast_info3[8] < $blast_info3[9])){
	  push @startq, $blast_info3[7];
	push @endq, $blast_info3[6];
	push @starth, $blast_info3[8];
	push @endh, $blast_info3[9];
	  }
	  elsif (($blast_info3[6] < $blast_info3[7]) && ($blast_info3[8] > $blast_info3[9])){
	  push @startq, $blast_info3[6];
	push @endq, $blast_info3[7];
	push @starth, $blast_info3[9];
	push @endh, $blast_info3[8];
	  }
	  else{
	  push @startq, $blast_info3[7];
	push @endq, $blast_info3[6];
	push @starth, $blast_info3[9];
	push @endh, $blast_info3[8];
	  }
	print OUTPUT2
"$blast_info3[0]\t$blast_info3[1]\t$blast_info3[2]\t$blast_info3[3]\t$blast_info3[4]\t$blast
_info3[5]\t$blast_info3[6]\t$blast_info3[7]\t$blast_info3[8]\t$blast_info3[9]\t$blast_info3
[10]\t$blast_info3[11]";
	  #start loop to read in all other lines add starts and ends to arrays
	while ($blast_table2 = <INPUT3>) {
		@blast_info2 = split('\t', $blast_table2);
```

```
                if (($blast_info2[6] < $blast_info2[7]) && ($blast_info2[8] <
$blast_info2[9])){
                push @startq, $blast_info2[6];
                push @endq, $blast_info2[7];
                push @starth, $blast_info2[8];
        push @endh, $blast_info2[9];
        }
        elsif (($blast_info2[6] > $blast_info2[7]) && ($blast_info2[8] < $blast_info2[9])){
        push @startq, $blast_info2[7];
        push @endq, $blast_info2[6];
        push @starth, $blast_info2[8];
        push @endh, $blast_info2[9];
        }
        elsif (($blast_info2[6] < $blast_info2[7]) && ($blast_info2[8] > $blast_info2[9])){
        push @startq, $blast_info2[6];
        push @endq, $blast_info2[7];
        push @starth, $blast_info2[9];
                push @endh, $blast_info2[8];
        }
                else{
        push @startq, $blast_info2[7];
        push @endq, $blast_info2[6];
        push @starth, $blast_info2[9];
        push @endh, $blast_info2[8];
        }
                $linecounter4 ++;
                $counter = 1; #this counter is used in order  to compare to all previous
hits
                $print_or_not = 0; #value added to if breaks conditions and means it is
not printed to new table
                while ($counter <= $linecounter4) {
                        if ($startq[$linecounter4] >= $startq[($linecounter4 - $counter)]
&& $endq[$linecounter4] <= $endq[($linecounter4 - $counter)] or
$startq[$linecounter4] < $endq[($linecounter4 - $counter)] && $endq[$linecounter4] >
$endq[($linecounter4 - $counter)] or $endq[$linecounter4] > $startq[($linecounter4 -
$counter)] && $startq[$linecounter4] < $startq[($linecounter4 - $counter)] or
$starth[$linecounter4] >= $starth[($linecounter4 - $counter)] && $endh[$linecounter4]
<= $endh[($linecounter4 - $counter)] or $starth[$linecounter4] < $endh[($linecounter4
- $counter)] && $endh[$linecounter4] > $endh[($linecounter4 - $counter)] or
$endh[$linecounter4] > $starth[($linecounter4 - $counter)] && $starth[$linecounter4] <
$starth[($linecounter4 - $counter)]) {
                        $print_or_not ++;
                        pop @starth; #remove last value from array if not printing as other
contig may come that doesn't overlap printed but does overlap unprinted
                  pop @endh;
                        pop @startq; #remove last value from array if not printing as other
contig may come that doesn't overlap printed but does overlap unprinted
                        pop @endq; # same as above
                        $linecounter4 = $linecounter4 -1; #same as above
                        last;
                        }
                        else {
                        $counter ++; #if doesn't break condition add one to counter so as
to compare against the next hit
```

```
                        }}
                        #print if it hasn't broken conditions and so = 0
                unless ($print_or_not != 0) {
                print OUTPUT2
"$blast_info2[0]\t$blast_info2[1]\t$blast_info2[2]\t$blast_info2[3]\t$blast_info2[4]\t$blast
_info2[5]\t$blast_info2[6]\t$blast_info2[7]\t$blast_info2[8]\t$blast_info2[9]\t$blast_info2
[10]\t$blast_info2[11]";
}}}
close INPUT3;

}

##########################################################################

sub remove_last_line_of_tab {

my $blast_redundancy_temp = $_[0];
open (INPUT2, "<$ARGV[0]_blast_redundancy_temp") or  die "could not open
blast_redundancy_temp for read\n";
my $first_line; my $blast_table; my @blast_info;
my $scaffold_to_keep;
$first_line = 0;
while ($blast_table = <INPUT2>) {
        @blast_info = split('\t', $blast_table);
        if ($first_line == 0) {
        $scaffold_to_keep = $blast_info[1];
        $first_line++;
        print TEMP3 "$blast_table";
        }
        elsif (($first_line > 0) && ($blast_info[1] eq $scaffold_to_keep)){
        print TEMP3 "$blast_table";
        }
        else {
        last;
        }
}
close INPUT2;
}
```

Appendix 15 – Script to do redundancy calculation

```
#! /usr/bin/perl -w
use strict;
use warnings;

# This script reads the sorted redundancy blast table (from redundancy blast table and
sort table) made # by blasting assembly contigs (redundancy filtered or not) against
reference  genome and calculates
# unique bases vs overlapping  # bases for each scaffold and then calculates overall
redundancy inthe # assembly
```

```
my $line; my @line_info; my $hit_start_total = 0; my $hit_end_total = 0; my
$overall_hit_bases = 0;
my $line2; my @line_info2; my @hit_start = (); my @hit_end = (); my @scaffold = ();
my $linecounter = 0; my $overall_redundant_bases = 0; my $backcount = 1;
my $redundancy_percentage;

# open the blast results table sorted
open (INPUTFILE, "<$ARGV[0]") or die "could not open sorted redundant blast results
table.\n";

#calculate overall number of bases hits cover
while ($line = <INPUTFILE>){
        @line_info =  split('\t', $line);
        $overall_hit_bases =  $overall_hit_bases + ($line_info[9] - $line_info[8]);
}

print "overall hit bases = $overall_hit_bases\n";
close INPUTFILE;
open (INPUTFILE, "<$ARGV[0]") or die "could not open sorted redundant blast results
table second time.\n";

while ($line2 = <INPUTFILE>){
        @line_info2 =  split('\t', $line2);
        push @hit_start, $line_info2[8];
      push @hit_end, $line_info2[9];
        push @scaffold, $line_info2[1];
        if ($linecounter == 1){
            if ($hit_start[$linecounter] <= $hit_end[($linecounter -1)]) {
                if ($hit_end[($linecounter)] >=  $hit_end[($linecounter -1)]) {
                        $overall_redundant_bases = $overall_redundant_bases +
($hit_end[($linecounter -1)] - $hit_start[$linecounter]);
                }
                elsif ($hit_end[($linecounter)] <  $hit_end[($linecounter -1)]) {
                        $overall_redundant_bases = $overall_redundant_bases +
($hit_end[($linecounter)] - $hit_start[$linecounter]);
                }
                else {
                        #dont add anything as non overlapping
                }
            }
        }
        elsif ($linecounter > 1){
                while ($scaffold[$linecounter] eq $scaffold[($linecounter -1)]){
                        while (($hit_start[$linecounter] <= $hit_end[($linecounter -
$backcount)]) && ($scaffold[$linecounter] eq $scaffold[($linecounter -$backcount)])) {
                                if ($hit_end[($linecounter)] >=  $hit_end[($linecounter -
$backcount)]) {
                                        $overall_redundant_bases =
$overall_redundant_bases + ($hit_end[($linecounter -$backcount)] -
$hit_start[$linecounter]);
                                        $backcount++;
                                }
                                elsif ($hit_end[($linecounter)] <  $hit_end[($linecounter -
$backcount)]) {
```

```
                                        $overall_redundant_bases =
$overall_redundant_bases + ($hit_end[($linecounter)] - $hit_start[$linecounter]);
                                        $backcount++;
                                }
                                else {
                                        #dont add anything as non overlapping
                                }
                        }
                $backcount = 1;
            last;
                }
        }
        $linecounter++;
}
print "overall redundant bases = $overall_redundant_bases\n";
$redundancy_percentage = (($overall_redundant_bases / $overall_hit_bases) * 100);
print "redundancy percentage = $redundancy_percentage\n";
```

Appendix 16 – Script to pull out sequence and reverse complement, for use with fixed differences.

```
#! /usr/bin/perl -w
use strict;
use warnings;

# give fasta reference and check positions and script returns sequence
# Use like: perl check_positions_in_fasta.pl MJ09_4015_ner_K40_8080-contigs.fa
3457188 323 333 | # cat >> Ner_denovo_Fixdiff_seq

## open fasta file
open (INPUTFILE, "<$ARGV[0]") or die "could not open fasta file.\n";

## input contig or scaffold name
my $ref = "$ARGV[1]";

#check start
my $start = "$ARGV[2]";

#check finish
my $finish = "$ARGV[3]";

open (TEMP, ">$ARGV[0]_fasta_tmp.$ref");
system("grep -A1 '>$ref' $ARGV[0] > $ARGV[0]_fasta_tmp.$ref");

close TEMP;

open (TEMP, "<$ARGV[0]_fasta_tmp.$ref");

my $info_line = <TEMP>;
```

```
my $seq = <TEMP>;

chomp $seq;
my @seq = split('', $seq);

for (my $i=$start; $i <= $finish; $i++) {
        print "@seq[($i -1)]";
}
print "\t";

for (my $i=$finish; $i >= $start; $i--) {
        if (@seq[($i -1)] eq 'A'){
                print "T";
        }
        elsif (@seq[($i -1)] eq 'C'){
            print "G";
      }
        elsif (@seq[($i -1)] eq 'G'){
            print "C";
      }
        elsif (@seq[($i -1)] eq 'T'){
            print "A";
      }
        elsif (@seq[($i -1)] eq 'R'){
            print "Y";
      }
      elsif (@seq[($i -1)] eq 'Y'){
            print "R";
      }
      elsif (@seq[($i -1)] eq 'S'){
            print "S";
      }
        elsif (@seq[($i -1)] eq 'W'){
            print "W";
      }
      elsif (@seq[($i -1)] eq 'K'){
            print "M";
      }
      elsif (@seq[($i -1)] eq 'M'){
            print "K";
      }
        elsif (@seq[($i -1)] eq 'N'){
            print "N";
      }
}

print "\n";

close TEMP;
close INPUTFILE;
system("rm $ARGV[0]_fasta_tmp.$ref");
```

Appendix 17 – Script for primer design, adds in missing data rows to VCF file from alignment reference

```
#! /usr/bin/perl -w
use strict;
use warnings;

# This script takes an a calls file from and a fasta scaffold file (for a single scaffold), and adds in rows
# filtered from the calls #file as Ns to make new calls file. This is useful for primer design.

# open the ref file
open (INPUTFILE1, "<$ARGV[0]") or die "could not open input ref scaffold fasta.\n";

# open the calls file
open (INPUTFILE2, "<$ARGV[1]") or die "could not open calls file for input.\n"; #be aware that you must have tabs delimting header line!!!!

my $counter = 1; my $line1; my $seq; my $header; my $line2; my @info; my $i; my @base_calls; my %called_bases;
my $value; my @ntaxa_info; my $scaff; my $key;

my $temp_output_file = "$ARGV[1]_all_bases_temp";
open (OUTPUTFILE, ">$temp_output_file");

$line1 = <INPUTFILE1>;
$seq = <INPUTFILE1>;

my $num_bases = length($seq);
my @ref_seq =  split('', $seq);

$header = <INPUTFILE2>;
print OUTPUTFILE "$header";
@ntaxa_info =  split('\t', $header);
my $header_length = @ntaxa_info;
my $ntaxa = ($header_length - 2);
print "$ntaxa\n";

while ($line2 = <INPUTFILE2>){
        @base_calls=();
        chomp $line2;
        @info =  split('\t', $line2);
        for ($i = 2; $i < $header_length; $i++) {
                push @base_calls, "\t$info[$i]";
        }
        $called_bases{$info[1]} = "@base_calls";
        $scaff = $info[0];
}

#print "$_ $called_bases{$_}\n" for (keys %called_bases);


while ($counter <= $num_bases){
```

```
        print OUTPUTFILE "$scaff\t";
        if (exists $called_bases{$counter}) {
                if (($key, $value) = each %called_bases){
                        print OUTPUTFILE "$key";
                        print OUTPUTFILE "$value";
                        print OUTPUTFILE "\n";
                }
        }
        else {
                print OUTPUTFILE "$counter";
                for ($i = 2; $i < $header_length; $i++) {
            print OUTPUTFILE "\tN";
             }
        print OUTPUTFILE "\n";
        }
$counter++;
}

system("sort -nk2 $temp_output_file > $ARGV[1]_all_bases");
system("rm $temp_output_file");
```

Appendix 18 – Script for primer design, adds reference calls to alignment file

```
#! /usr/bin/perl -w
use strict;
use warnings;

# Script for primer design, adds reference calls to alignment file


my $line; my $i; my @value; my %genome; my %scaffold_lengths;


open CALLS_FILE, "<$ARGV[0]" or die "could not open calls file (first argument).\n";

open REF_FASTA, "<$ARGV[1]" or die "could not open reference fasta file (second
argument). NEEDS TO BE A ONE LINE FASTA FORMAT.\n";

my $ref_species = $ARGV[2] or die "provide reference species (third argument. No
spaces or weird characters.\n";

my $output = $ref_species . '_' . $ARGV[0];

open OUTPUT, ">$output";

# Read reference into a hash
while ($line = <REF_FASTA>) {
       chomp($line);
         my $scaffold = substr($line, 1);
         $line = <REF_FASTA>;
         chomp($line);
         my $length = length($line);
```

```
		$scaffold_lengths{$scaffold} = $length;
		$genome{$scaffold} = $line;
}

# Read and print the header line
$line = <CALLS_FILE>;
chomp($line);
print OUTPUT "$line\t$ref_species\n";

while ($line = <CALLS_FILE>) {
		chomp($line);
		@value = split(' ', $line);

		my $ref_scaffold = $value[0];
		my $ref_position = $value[1];

		if (exists $genome{$ref_scaffold}) {
			if ($ref_position <= $scaffold_lengths{$ref_scaffold}) {
				my $ref_position2 = $ref_position - 1;
				my $ref_base = substr($genome{$ref_scaffold}, $ref_position2, 1);
				print OUTPUT "$line\t$ref_base\n";
			}
			else {
				print "Requested scaffold position $ref_scaffold $ref_position is
greater than the actual scaffold length of $scaffold_lengths{$ref_scaffold}\n";
				die;
			}
		}
		else {
			print "Unknown genome scaffold $ref_scaffold encountered\n";
			die;
		}
}

close CALLS_FILE;
close REF_FASTA;
close OUTPUT;
```

Appendix 19 – Script to align de novo contigs to given reference

```
#! /usr/bin/perl -w
use strict;
use warnings;

# script to make a fasta file for those contigs not aligned by needle, instead uses blast to
align them
# (reverese translates those that it is needed for) input is the fasta file of fixed difference
contigs

open (INPUTFILE, "<$ARGV[0]") or die "could not open fasta file of contigs.\n";
```

```
my $output_file1 = "$ARGV[0].blasted_fasta_alignment";
open (OUTPUT, ">$output_file1");

my $blast_ref = $ARGV[1];

my $line1; my $line2; my @contig_info; my  $contig_name;
my $loop_num = 0; my $line_num = 0; my $crap_line; my @query_info; my
$query_start; my $subjt_line; my @subjt_info; my $subjt_start; my $num_of_hyphens;
my $info_line; my $contig; my $second_subjt; my @subjt_info2; my $subjt_start2 ; my
$print_num = 0; my $ref_size_line; my @ref_size_info; my $ref_size;
my $line_num2 = 0; my $contig_length; my $print_num2 = 0; my $num_of_hyphens2; my
$hyph_start; my $rev_trans_num; my @bases;

while ($line1 = <INPUTFILE>) {
        $line2 = <INPUTFILE>;
        @contig_info = split(' ', $line1);
      $contig_name = substr $contig_info[0],1;
        open (TEMP, ">$ARGV[0]_contig_tmp_$contig_name");
        system("grep -A1 '$contig_info[0]' $ARGV[0] >
$ARGV[0]_contig_tmp_$contig_name");
        close TEMP;
        open (TEMP, "<$ARGV[0]_contig_tmp_$contig_name");
        $info_line = <TEMP>;
        $contig = <TEMP>;
        chomp $contig;
        system("/usr/local/src/ncbi-blast-2.2.27+/bin/blastn -task blastn -db $blast_ref -
query $ARGV[0]_contig_tmp_$contig_name -outfmt 0 > $ARGV[0]_$contig_name");
      system("rm $ARGV[0]_contig_tmp_$contig_name");

        #open blast output and process
        open (INPUTFILE4, "<$ARGV[0]_$contig_name") or die "could not open blast
output.\n";
        while ($line_num <= 30) {
                $crap_line = <INPUTFILE4>;
                if ($line_num == 24) {
                        $ref_size_line = <INPUTFILE4>;
                        @ref_size_info = split('=', $ref_size_line);
                  $ref_size = $ref_size_info[1];
                        $line_num++;
                }
                elsif ($line_num == 30) {
                        @query_info = split(' ', $crap_line);
                        $query_start = $query_info[1];
                        $crap_line = <INPUTFILE4>;
                        $subjt_line = <INPUTFILE4>;
                        @subjt_info = split(' ', $subjt_line);
                      $subjt_start = $subjt_info[1];
                        $crap_line = <INPUTFILE4>;
                        $crap_line = <INPUTFILE4>;
                        $crap_line = <INPUTFILE4>;
                        $second_subjt = <INPUTFILE4>;
                        @subjt_info2 = split(' ', $second_subjt);
                $subjt_start2 = $subjt_info2[1];
```

```
                        if ($subjt_start2 eq '=') {
                                system("rm $ARGV[0]_$contig_name");
                                last;
                        }
                        elsif ($subjt_start2 eq 'K') {
                        system("rm $ARGV[0]_$contig_name");
                                last;
                        }
                        elsif ($subjt_start2 > $subjt_start) {
                                $num_of_hyphens = ($subjt_start - $query_start);
                                print OUTPUT ">$contig_name\n";
                                while ($print_num < $num_of_hyphens) {
                                        print OUTPUT "-";
                                        $print_num++;
                                }
                                print OUTPUT "$contig";
                                $contig_length = length($contig);
                                $hyph_start = ($contig_length + $num_of_hyphens);
                                $num_of_hyphens2 = ($ref_size - $hyph_start);

                                system("rm $ARGV[0]_$contig_name");
                        }
                        elsif ($subjt_start2 < $subjt_start) {
                                print OUTPUT ">$contig_name\n";
                                $contig_length = length($contig);
                                $num_of_hyphens = (($subjt_start + $query_start) -
$contig_length);
                                while ($print_num < $num_of_hyphens) {
                              print OUTPUT "-";
                              $print_num++;
                                }
                                @bases = split(//, $contig);    #now print from last base and
translate till 0
                                $rev_trans_num = ($contig_length - 1);
                                while ($rev_trans_num >= 0  ) {
                                        if ($bases[$rev_trans_num] eq 'A') {
                                                print OUTPUT "T";
                                        }
                                        elsif ($bases[$rev_trans_num] eq 'T') {
                                    print OUTPUT "A";
                              }
                                        elsif ($bases[$rev_trans_num] eq 'C') {
                                    print OUTPUT "G";
                              }
                                        elsif ($bases[$rev_trans_num] eq 'G') {
                                    print OUTPUT "C";
                              }
                                        else {
                                                print OUTPUT "N";
                                        }
                                        $rev_trans_num--;
                                }
                                $hyph_start = ($contig_length + $num_of_hyphens);
                        $num_of_hyphens2 = ($ref_size - $hyph_start);
```

```
                        system("rm $ARGV[0]_$contig_name");
                }
                while ($print_num2 < $num_of_hyphens2) {
                        print OUTPUT "-";
                        $print_num2++;
                }
                        print OUTPUT "\n";
                        last;
        }
        else {
                $line_num++;
                $print_num2 = 0;
                $print_num = 0;
                $num_of_hyphens2 = 0;
                $num_of_hyphens = 0;
        }
    }
    $line_num = 0
}
```

Appendix 20 – Script to filter VCF table by quality thresholds, and arranges to filtered SNP and indel files

```
#! /usr/bin/perl -w
use strict;
use warnings;
use Getopt::Std;

# script takes a file made from GATKs VariantsToTable program. Needs to take arguments -F
# CHROM -F POS -F REF -F ALT -F # QUAL -F MQ \ -GF GQ -GF DP -GF GT\
# this means table heading should be CHROM POS REF ALT QUAL MQ (then for each sample)
# GQ DP GT outputs seperate calls files of snps and indels

our ($opt_q, $opt_g, $opt_h, $opt_l, $opt_i, $opt_m);

getopt("qghlmi");

#options
my $SNPqual;  if ($opt_q) {$SNPqual = $opt_q;} else {die "provide SNP quality threshold using -q option\n";}
my @qual; if ($opt_g) {@qual = split(":", $opt_g);} else {die "provide genotype quality threshold using -g option\n";}
my @hicover; if ($opt_h) {@hicover = split(":", $opt_h);} else {die "provide maximum coverage threshold with -h option\n";}
my @lowcover; if ($opt_l) {@lowcover = split(":", $opt_l);} else {die "provide minimum coverage threshold with -l option\n";}
my $input_file; if ($opt_i) {$input_file = $opt_i;} else {die "provide vcf input with -i option\n";}
```

```
my $min_MQ; if ($opt_m) {$min_MQ = $opt_m;} else {die "provide minimum mapping
quality threshold with -m option\n";}

my $header_line; my @header_info; my $short; my $i; my $ntaxa; my
@counthomREF_snp; my @counthomALT_snp; my @counthet_snp; my @countN;
my @counthicov_snp; my @countlowcov_snp; my @countweird; my @site_info; my
$line; my $ref_length; my $calls_length; my @genotype_calls;
my @consensus; my $count; my @calls_info; my $calls_array_length; my
$possible_indel_length; my $j; my $k; my $l; my $multiallele = 0;
my @countlowgenoqual_snp; my @countlowgenoqual_indel; my @counthomREF_indel;
my @count_indel_het; my @counthomALT_indel;
my @count_indel_N; my @counthicov_indel; my @countlowcov_indel; my @names; my
@count_other_snp; my @count_other_indel;
my $indel_decider = 0; my $length_checker;

# open GATK results table from .VCF file
open INPUTFILE, "<$input_file" or die "could not open input GATK results table.\n";

my $log_file = $input_file . '_FILTER_LOG';
open (LOGFILE, ">$log_file");

my $output_file = $input_file . '_filtered_snps';
open (OUTPUT1, ">$output_file");

my $output_file2 = $input_file . '_filtered_indels';
open (OUTPUT2, ">$output_file2");

#my $output_file3 = $input_file . '_weirdtest';
#open (OUTPUT3, ">$output_file3");

#sort out header and count ntaxa
for ($header_line = <INPUTFILE>) {
        @header_info =  split('\t', $header_line);
        print OUTPUT1 "$header_info[0]\t$header_info[1]";
        print OUTPUT2 "$header_info[0]\t$header_info[1]";
        $ntaxa = ((@header_info - 6)/3);
        for ($i = 6; $i < (($ntaxa * 3)+ 6); $i += 3) {
                $short = substr($header_info[$i], 0, -3);
                print OUTPUT1 "\t$short";
                print OUTPUT2 "\t$short";
                push @names, $short;
        }
        print OUTPUT1 "\n";
        print OUTPUT2 "\n";
}

#print "@names\n";

my $countlowqual = 0;
my $countlowMQ = 0;

#set up counters for logfile
for ($i = 0; $i < $ntaxa ; $i++) {
```

```
        $counthomREF_snp[$i] = 0; $counthomALT_snp[$i] = 0; $counthet_snp[$i] = 0;
$countN[$i] = 0;
        $counthicov_snp[$i] = 0; $countlowcov_snp[$i] = 0; $countweird[$i] = 0;
        $count_indel_het[$i] = 0; $count_indel_N[$i] = 0; $counthicov_indel[$i] = 0;
$countlowcov_indel[$i] = 0;
        $countlowgenoqual_snp[$i] = 0; $countlowgenoqual_indel[$i] = 0;
$counthomREF_indel[$i] = 0; $counthomALT_indel[$i] = 0;
        $count_other_snp[$i] = 0; $count_other_indel[$i] = 0;
}

#print "ntaxa = $ntaxa";

# check that number of taxa in vcf matches filtering parameters provided
if ($ntaxa != @qual || $ntaxa != @lowcover || $ntaxa != @hicover) {
      die "mismatch between expected and actual number of taxa $ntaxa\n"
}


while ($line = <INPUTFILE>) {
        $multiallele = 0;
        $j = 6;
        $k = 7;
        $l = 8;
        $indel_decider = 0;
        @site_info =  split('\t', $line);
        $ref_length = length($site_info[2]);
      $calls_length = length($site_info[3]);
        #remove sites with poor mapping or SNPquality
        if ($site_info[4] < $SNPqual) {
                $countlowqual++;
        }
        elsif ($site_info[5] < $min_MQ) {
                $countlowMQ++;
        }
        else {  #sites that have a ref base then '.' e.g A .
                if (($site_info[3] eq '.') && ($ref_length == 1)){
                        print OUTPUT1 "$site_info[0]\t$site_info[1]";
                  for ($i = 0; $i < $ntaxa; $i++) {
                                if ($site_info[($k)] eq 'NA'){
                                print OUTPUT1 "\tN";
                                $count_other_snp[$i]++;
                        }
                                elsif (($lowcover[$i] <= $site_info[($k)]) &&
($site_info[($k)] <= $hicover[$i])) {
                                        chomp  $site_info[($l)];
                                        @genotype_calls = split('/', $site_info[($l)]);
                                if ($genotype_calls[0] eq $genotype_calls[1]) {
                                    if ($genotype_calls[0] eq ($site_info[2] or '.')){
                                                print OUTPUT1
"\t$genotype_calls[0]";
                                                $counthomREF_snp[$i]++;
                                        }
                                        else {
```

```
                                                print OUTPUT1
"\t$genotype_calls[0]";
                                                $counthomALT_snp[$i]++;
                                        }
                                }
                                else {
                                        #do heterozygosity check
                                        $consensus[$i] = HetBase($genotype_calls[0],
$genotype_calls[1]);
                                        print OUTPUT1 "\t$consensus[$i]";
                                                $counthet_snp[$i]++;
                                }
                        }
                          elsif ($lowcover[$i] > $site_info[($k)]){
                                print OUTPUT1 "\tN";
                                        $countlowcov_snp[$i]++;
                                        $countN[$i]++;
                          }
                          elsif ($site_info[($k)] > $hicover[$i]){
                                        print OUTPUT1 "\tN";
                                $counthicov_snp[$i]++;
                                $countN[$i]++;
                          }
                          else {
                                        print OUTPUT1 "\tN";
                                        $countweird[$i]++;
                          }
                    $j+=3;
                    $k+=3;
                    $l+=3;
                    }
              print OUTPUT1 "\n";
              }    #sites that have a ref indel then '.' e.g AAA .
                   elsif (($site_info[3] eq '.') && ($ref_length > 1)){
                            print OUTPUT2 "$site_info[0]\t$site_info[1]";
                            for ($i = 0; $i < $ntaxa; $i++) {
                            if ($site_info[($k)] eq 'NA'){
                                print OUTPUT2 "\tN";
                                $count_other_indel[$i]++;
                            }
                              elsif (($lowcover[$i] <= $site_info[($k)]) &&
($site_info[($k)] <= $hicover[$i])) {
                              chomp  $site_info[($l)];
                                        @genotype_calls = split('/', $site_info[($l)]);
                                if ($genotype_calls[0] eq $genotype_calls[1]) {
                                        if ($genotype_calls[0] eq ($site_info[2] or '.')){
                                          print OUTPUT2 "\t$genotype_calls[0]";
                                          $counthomREF_indel[$i]++;
                                    }
                                            else {
                                          print OUTPUT2 "\t$genotype_calls[0]";
                                          $counthomALT_indel[$i]++;
                                    }
                                }
```

```
                        else {
                            print OUTPUT2 "\t$site_info[($l)]";
                                $count_indel_het[$i]++;
                        }
                    }
                       elsif ($lowcover[$i] > $site_info[($k)]){
                      print OUTPUT2 "\tN";
                        $countlowcov_indel[$i]++;
                        $count_indel_N[$i]++;
                    }
                    elsif ($site_info[($k)] > $hicover[$i]){
                      print OUTPUT2 "\tN";
                        $counthicov_indel[$i]++;
                        $count_indel_N[$i]++;
                    }
                       else {
                        print OUTPUT2 "\tN";
                            $countweird[$i]++;
                    }
                     $j+=3;
                  $k+=3;
                  $l+=3;
                  }
                print OUTPUT2 "\n";
                }       #sites that have a ref base then a sigle alt allele  e.g A C
                elsif (($site_info[3] ne '.') && ($ref_length == 1) && ($calls_length == 1)) {
                        #then SNP and Do SNP
                        print OUTPUT1 "$site_info[0]\t$site_info[1]";
                        for ($i = 0; $i < $ntaxa; $i++) {
                            if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                        print OUTPUT1 "\tN";
                        $count_other_snp[$i]++;
                    }
                            elsif (($site_info[($j)] >= $qual[$i]) && (($lowcover[$i] <=
$site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                    chomp  $site_info[($l)];
                                    @genotype_calls = split('/', $site_info[($l)]);
                                    if ($genotype_calls[0] eq $genotype_calls[1]) {
                                            if ($genotype_calls[0] eq ($site_info[2] or
'.')){
                                      print OUTPUT1 "\t$genotype_calls[0]";
                                      $counthomREF_snp[$i]++;
                                }
                                            else {
                                      print OUTPUT1 "\t$genotype_calls[0]";
                                      $counthomALT_snp[$i]++;
                                }
                                    }
                                    else {
                                            #do heterozygosity check
                                            $consensus[$i] =
HetBase($genotype_calls[0], $genotype_calls[1]);
                                            print OUTPUT1 "\t$consensus[$i]";
                                            $counthet_snp[$i]++;
```

```
                                        }
                                }
                                elsif ($lowcover[$i] > $site_info[($k)]){
                                print OUTPUT1 "\tN";
                                $countlowcov_snp[$i]++;
                                $countN[$i]++;
                        }
                        elsif ($site_info[($k)] > $hicover[$i]){
                                print OUTPUT1 "\tN";
                                $counthicov_snp[$i]++;
                                $countN[$i]++;
                        }
                                elsif ($site_info[($j)] < $qual[$i]){
                                        print OUTPUT1 "\tN";
                                        $countlowgenoqual_snp[$i]++;
                                        $countN[$i]++;
                                }
                                else {
                                        print OUTPUT1 "\tN";
                                $countweird[$i]++;
                        }
                        $j+=3;
                        $k+=3;
                        $l+=3;
                        }
                print OUTPUT1 "\n";
                }       #sites that have a ref base then an alt indel or multiple alt allele  e.g
A C,T or A AAAA or A ATA,AAA
                elsif (($site_info[3] ne '.') && ($ref_length == 1) && ($calls_length > 1)) {
                        @calls_info = split(',', $site_info[3]);
                  $calls_array_length = @calls_info;
                  $count = 0;
                        while ($count < $calls_array_length){
                        $possible_indel_length = length($calls_info[$count]);
                                if ($possible_indel_length > 1){        #sites that have a ref
base then an alt indel e.g A AAA or A ATA,AAA
                                        $multiallele = 1;
                                        #then indel
                                        print OUTPUT2 "$site_info[0]\t$site_info[1]";
                                        for ($i = 0; $i < $ntaxa; $i++) {
                                                #print "$site_info[($j)]\t$site_info[($k)]\n";
                                                if (($site_info[($k)] or $site_info[($j)]) eq
'NA'){
                                                print OUTPUT2 "\tN";
                                        $count_other_indel[$i]++;
                                }
                                                elsif (($site_info[($j)] >= $qual[$i]) &&
(($lowcover[$i] <= $site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                                        chomp  $site_info[($l)];
                                                        @genotype_calls = split('/',
$site_info[($l)]);
                                                        if ($genotype_calls[0] eq
$genotype_calls[1]) {
```

```
                                                        if ($genotype_calls[0] eq
($site_info[2] or '.')){
                                                    print OUTPUT2 "\t$genotype_calls[0]";
                                                      $counthomREF_indel[$i]++;
                                                }
                                                        else {
                                                      print OUTPUT2
"\t$genotype_calls[0]";
                                                      $counthomALT_indel[$i]++;
                                                }
                                                      }
                                                      else {
                                                            print OUTPUT2
"\t$site_info[($l)]";
                                                            $count_indel_het[$i]++;
                                                      }
                                                }
                                                      elsif ($lowcover[$i] > $site_info[($k)]){
                                          print OUTPUT2 "\tN";
                                          $countlowcov_indel[$i]++;
                                          $count_indel_N[$i]++;
                                    }
                                    elsif ($site_info[($k)] > $hicover[$i]){
                                          print OUTPUT2 "\tN";
                                          $counthicov_indel[$i]++;
                                          $count_indel_N[$i]++;
                                    }
                                    elsif ($site_info[($j)] < $qual[$i]){
                                          print OUTPUT2 "\tN";
                                          $countlowgenoqual_indel[$i]++;
                                          $count_indel_N[$i]++;
                                    }
                                                else {
                                                      print OUTPUT2 "\tN";
                                                $countweird[$i]++;
                                    }
                                          $j+=3;
                                      $k+=3;
                                          $l+=3;
                                          }
                                    print OUTPUT2 "\n";
                                    last;
                                    }
                                    #sites that have a ref base and multiple alt allele  e.g A C,T
                                    elsif (($count == ($calls_array_length -1)) && ($multiallele
== 0)){
                                                #then SNP and Do SNP
                              print OUTPUT1 "$site_info[0]\t$site_info[1]";
                                          for ($i = 0; $i < $ntaxa; $i++) {
                                          if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                                          print OUTPUT1 "\tN";
                                                $count_other_snp[$i]++;
                                    }
```

```
                                        elsif (($site_info[($j)] >= $qual[$i]) &&
(($lowcover[$i] <= $site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                        chomp  $site_info[($l)];
                                                @genotype_calls = split('/',
$site_info[($l)]);
                                        if ($genotype_calls[0] eq $genotype_calls[1])
{
                                                if ($genotype_calls[0] eq ($site_info[2]
or '.')){
                                                print OUTPUT1
"\t$genotype_calls[0]";
                                                $counthomREF_snp[$i]++;
                                        }
                                                        else {
                                                print OUTPUT1
"\t$genotype_calls[0]";
                                                $counthomALT_snp[$i]++;
                                        }
                                        }
                                        else {
                                                #do heterozygosity check
                                        $consensus[$i] =
HetBase($genotype_calls[0], $genotype_calls[1]);
                                        print OUTPUT1 "\t$consensus[$i]";
                                                        $counthet_snp[$i]++;
                                                }
                                }
                                        elsif ($lowcover[$i] > $site_info[($k)]){
                                      print OUTPUT1 "\tN";
                                $countlowcov_snp[$i]++;
                                $countN[$i]++;
                        }
                        elsif ($site_info[($k)] > $hicover[$i]){
                                print OUTPUT1 "\tN";
                                $counthicov_snp[$i]++;
                                $countN[$i]++;
                        }
                        elsif ($site_info[($j)] < $qual[$i]){
                                print OUTPUT1 "\tN";
                                $countlowgenoqual_snp[$i]++;
                                $countN[$i]++;
                        }
                                        else {
                                                print OUTPUT1 "\tN";
                                      $countweird[$i]++;
                        }
                                $j+=3;
                $k+=3;
                        $l+=3;
                        }
                        print OUTPUT1 "\n";
                        }
                $count++;
                }
```

```
                }
                #in case reference is indel and within alt is an indel e.g AAA ATA or
multisnp e.g AAA A,T,C
                elsif (($site_info[3] ne '.') && ($ref_length > 1) && ($calls_length > 1)){
                        @calls_info = split(',', $site_info[3]);
                  $calls_array_length = @calls_info;
                  $count = 0;
                        for ($i = 8; $i <= (($ntaxa*3) + 6); $i+=3) {
                        chomp  $site_info[($i)];
                        $length_checker = length($site_info[$i]);
                        if ($length_checker > 3) {
                                $indel_decider++;
                        }
                  }
                  if ($indel_decider > 0) {
                        #sites that have a ref indel and then an alt indel e.g AA AAA or AA
ATA,AAA
                        #then indel
                        print OUTPUT2 "$site_info[0]\t$site_info[1]";
                        for ($i = 0; $i < $ntaxa; $i++) {
                                if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                                        print OUTPUT2 "\tN";
                                  $count_other_indel[$i]++;
                        }
                                        elsif (($site_info[($j)] >= $qual[$i]) &&
(($lowcover[$i] <= $site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                                chomp  $site_info[($l)];
                                        @genotype_calls = split('/', $site_info[($l)]);
                                        if ($genotype_calls[0] eq $genotype_calls[1]) {
                                             if ($genotype_calls[0] eq ($site_info[2] or '.')){
                                                        print OUTPUT2 "\t$genotype_calls[0]";
                                                        $counthomREF_indel[$i]++;
                                                 }
                                                                else {
                                                        print OUTPUT2 "\t$genotype_calls[0]";
                                                        $counthomALT_indel[$i]++;
                                                 }
                                                        }
                                        else {
                                                 print OUTPUT2 "\t$site_info[($l)]";
                                                                $count_indel_het[$i]++;
                                        }
                                }
                                elsif ($lowcover[$i] > $site_info[($k)]){
                                  print OUTPUT2 "\tN";
                                        $countlowcov_indel[$i]++;
                                        $count_indel_N[$i]++;
                                }
                                elsif ($site_info[($k)] > $hicover[$i]){
                                  print OUTPUT2 "\tN";
                                        $counthicov_indel[$i]++;
                                        $count_indel_N[$i]++;
                                }
                                elsif ($site_info[($j)] < $qual[$i]){
```

```
                                print OUTPUT2 "\tN";
                                $countlowgenoqual_indel[$i]++;
                                $count_indel_N[$i]++;
                        }
                                else {
                                        print OUTPUT2 "\tN";
                                  $countweird[$i]++;
                        }
                                $j+=3;
                        $k+=3;
                        $l+=3;
                        }
                        print OUTPUT2 "\n";
                        }
#is this indel?  Sites that have a ref indel and multiple alt allele  e.g AAA C,T
                        elsif ($indel_decider == 0) {
                                #then SNP and Do SNP
                                print OUTPUT1 "$site_info[0]\t$site_info[1]";
                         for ($i = 0; $i < $ntaxa; $i++) {
                                if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                                        print OUTPUT1 "\tN";
                                        $count_other_snp[$i]++;
                                }
                                        elsif (($site_info[($j)] >= $qual[$i]) &&
(($lowcover[$i] <= $site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                        chomp  $site_info[($l)];
                                          @genotype_calls = split('/', $site_info[($l)]);
                                          if ($genotype_calls[0] eq $genotype_calls[1]) {
                                                if ($genotype_calls[0] eq ($site_info[2] or
'.')){
                                            print OUTPUT1 "\t$genotype_calls[0]";
                                                $counthomREF_snp[$i]++;
                                        }
                                                else {
                                                print OUTPUT1 "\t$genotype_calls[0]";
                                                $counthomALT_snp[$i]++;
                                        }
                                           }
                                           else {
                                #do heterozygosity check
                                                $consensus[$i] =
HetBase($genotype_calls[0], $genotype_calls[1]);
                                        print OUTPUT1 "\t$consensus[$i]";
                                                $counthet_snp[$i]++;
                                           }
                                }
                                        elsif ($lowcover[$i] > $site_info[($k)]){
                                print OUTPUT1 "\tN";
                                        $countlowcov_snp[$i]++;
                                        $countN[$i]++;
                                }
                                elsif ($site_info[($k)] > $hicover[$i]){
                                        print OUTPUT1 "\tN";
                                        $counthicov_snp[$i]++;
```

```
                                $countN[$i]++;
                        }
                        elsif ($site_info[($j)] < $qual[$i]){
                                print OUTPUT1 "\tN";
                                $countlowgenoqual_snp[$i]++;
                                $countN[$i]++;
                        }
                                else {
                                        print OUTPUT1 "\tN";
                                $countweird[$i]++;
                        }
                          $j+=3;
                                $k+=3;
                          $l+=3;
                    }
                          print OUTPUT1 "\n";
                }
                }
                #in case reference is indel but within alt is a base e.g AAA A
                elsif (($site_info[3] ne '.') && ($ref_length > 1) && ($calls_length == 1)){
                        for ($i = 8; $i <= (($ntaxa*3) + 6); $i+=3) {
                                chomp  $site_info[($i)];
                                $length_checker = length($site_info[$i]);
                                if ($length_checker > 3) {
                                        $indel_decider++;
                                }
                  }
                        if ($indel_decider == 0) {
                                #then snp
                     print OUTPUT1 "$site_info[0]\t$site_info[1]";
                       for ($i = 0; $i < $ntaxa; $i++) {
                                        if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                                print OUTPUT1 "\tN";
                                $count_other_snp[$i]++;
                                }
                                        elsif (($site_info[($j)] >= $qual[$i]) &&
(($lowcover[$i] <= $site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                                chomp  $site_info[($l)];
                                                @genotype_calls = split('/', $site_info[($l)]);
                                                if ($genotype_calls[0] eq $genotype_calls[1])
{
                                                        if ($genotype_calls[0] eq
($site_info[2] or '.')){
                                        print OUTPUT1 "\t$genotype_calls[0]";
                                          $counthomREF_snp[$i]++;
                                }
                                                        else {
                                        print OUTPUT1 "\t$genotype_calls[0]";
                                          $counthomALT_snp[$i]++;
                                }
                                                }
                                                else {
                                                        #do heterozygosity check
```

```
                                                $consensus[$i] =
HetBase($genotype_calls[0], $genotype_calls[1]);
                                    print OUTPUT1 "\t$consensus[$i]";
                                    $counthet_snp[$i]++;
                                                }
                                }
                                        elsif ($lowcover[$i] > $site_info[($k)]){
                                    print OUTPUT1 "\tN";
                                    $countlowcov_snp[$i]++;
                                    $countN[$i]++;
                                }
                            elsif ($site_info[($k)] > $hicover[$i]){
                                    print OUTPUT1 "\tN";
                                    $counthicov_snp[$i]++;
                                    $countN[$i]++;
                            }
                            elsif ($site_info[($j)] < $qual[$i]){
                                            print OUTPUT1 "\tN";
                                    $countlowgenoqual_snp[$i]++;
                                    $countN[$i]++;
                            }
                                            else {
                                                    print OUTPUT1 "\tN";
                                      $countweird[$i]++;
                            }
                            $j+=3;
                            $k+=3;
                            $l+=3;
                            }
                        print OUTPUT1 "\n";
                            }
                            elsif ($indel_decider > 0) {
                                      #then indel
                                      print OUTPUT2 "$site_info[0]\t$site_info[1]";
                                      for ($i = 0; $i < $ntaxa; $i++) {
                                    if (($site_info[($k)] or $site_info[($j)]) eq 'NA'){
                                    print OUTPUT2 "\tN";
                                        $count_other_indel[$i]++;
                                    }
                                    elsif (($site_info[($j)] >= $qual[$i]) && (($lowcover[$i] <=
$site_info[($k)]) && ($site_info[($k)] <= $hicover[$i]))) {
                                        chomp  $site_info[($l)];
                                        @genotype_calls = split('/', $site_info[($l)]);
                                        if ($genotype_calls[0] eq $genotype_calls[1]) {
                                            if ($genotype_calls[0] eq ($site_info[2] or '.')){
                                                    print OUTPUT2 "\t$genotype_calls[0]";
                                                    $counthomREF_indel[$i]++;
                                             }
                                             else {
                                                    print OUTPUT2 "\t$genotype_calls[0]";
                                                    $counthomALT_indel[$i]++;
                                             }
                                        }
                                        else {
```

```
                                        print OUTPUT2 "\t$site_info[($l)]";
                                        $count_indel_het[$i]++;
                                    }
                                }
                                elsif ($lowcover[$i] > $site_info[($k)]){
                                   print OUTPUT2 "\tN";
                                        $countlowcov_indel[$i]++;
                                        $count_indel_N[$i]++;
                                }
                                elsif ($site_info[($k)] > $hicover[$i]){
                                                print OUTPUT2 "\tN";
                                        $counthicov_indel[$i]++;
                                        $count_indel_N[$i]++;
                                }
                                elsif ($site_info[($j)] < $qual[$i]){
                                   print OUTPUT2 "\tN";
                                        $countlowgenoqual_indel[$i]++;
                                        $count_indel_N[$i]++;
                                }
                                else {
                                        print OUTPUT2 "\tN";
                                       $countweird[$i]++;
                                }
                            $j+=3;
                            $k+=3;
                            $l+=3;
                            }
                             print OUTPUT2 "\n";
                      }
                    }
            }
}

# Print summary to screen

print LOGFILE "Sites removed for poor mapping quality = $countlowMQ\n";
print LOGFILE "Sites removed for low SNP quality = $countlowqual\n\n\n";

for ($i = 0; $i < $ntaxa ; $i++) {
          print LOGFILE "$names[$i]\n\nSNP POSITIONS\n";
        print LOGFILE "No coverage positions = $countN[$i]\n";
          print LOGFILE "Positions with NA depth or GQ  = $count_other_snp[$i]\n";
        print LOGFILE "Low GQ positions = $countlowgenoqual_snp[$i]\n";
        print LOGFILE "Positions with too high coverage = $counthicov_snp[$i]\n";
        print LOGFILE "Positions with too low coverage = $countlowcov_snp[$i]\n";
        print LOGFILE "Homozygous REF positions = $counthomREF_snp[$i]\n";
        print LOGFILE "Homozygous ALT positions = $counthomALT_snp[$i]\n";
        print LOGFILE "Heterozygous positions = $counthet_snp[$i]\n\nINDEL
POSITIONS\n";
        print LOGFILE "No coverage positions = $count_indel_N[$i]\n";
          print LOGFILE "Positions with NA depth or GQ  = $count_other_indel[$i]\n";
        print LOGFILE "Low GQ positions = $countlowgenoqual_indel[$i]\n";
        print LOGFILE "Positions with too high coverage = $counthicov_indel[$i]\n";
        print LOGFILE "Positions with too low coverage = $countlowcov_indel[$i]\n";
```

```
        print LOGFILE "Homozygous REF positions = $counthomREF_indel[$i]\n";
        print LOGFILE "Homozygous ALT positions = $counthomALT_indel[$i]\n";
        print LOGFILE "Heterozygous positions = $count_indel_het[$i]\n\n";
          print LOGFILE "Weird positions (should be zero) = $countweird[$i]\n\n";
}



sub HetBase {
        my($retval); my($key); my($base) = $_[0] . $_[1];
        my (%ambigbases) = (AC => 'M',CA => 'M',CT => 'Y',TC => 'Y',GA => 'R',AG =>
'R',GT => 'K',TG => 'K',CG => 'S',GC => 'S',AT => 'W',TA => 'W');
        foreach $key (keys(%ambigbases)) {
                if ($base eq $key) {
                        $retval = $ambigbases{$key};
                        last;
                }
                else {
                        $retval = 'N';
                }
        }
        return $retval;
}
```

Appendix 21 – Script removes SNPs with missing data above threshold from calls file and estimates missing data per sample

```
#! /usr/bin/perl -w
use strict;
use warnings;

#script removes lines with  more Ns than threhold proportion of missing data to allow,
pipe output to logfile

open (INPUTFILE, "<$ARGV[0]") or die "could not open snp calls file.\n";

my $line; my $line1; my @line_stuff; my $ntaxa; my $N_num = 0; my $num = 2; my
$header;
my @Ns_per_sample; my @N_for_site; my $i; my $site_num = 0; my @prop_of_Ns;

my $threshold = ($ARGV[1]) or die "please enter threhold as proportion or missing data
to allow.\n";

my $output_file1 = "$ARGV[0]_no_Ns_thresh$threshold";
open (OUTPUT, ">$output_file1");


my $N = 'N';

$header = <INPUTFILE>;
```

```
chomp $header;
my @header_stuff = split'\t', $header;
my $length = @header_stuff;
$ntaxa = ($length - 2);

for ($i = 0; $i < $ntaxa; ++$i){
	$Ns_per_sample[$i] = 0;
}

for ($i = 0; $i < $ntaxa; ++$i){
	$N_for_site[$i] = 0;
}


#print "$ntaxa\n";
print OUTPUT "$header\n";

while ($line1 = <INPUTFILE>) {
	chomp $line1;
	@line_stuff = split'\t', $line1;
	while ($num < ($ntaxa + 2)) {
		if ($line_stuff[$num] eq $N) {
			$N_num++;
			$N_for_site[($num -2)]++;
			$num++;
		}
		else{
			$num++;
		}
	}
	if (($N_num / $ntaxa) <= $threshold ){
		print OUTPUT "$line1\n";
		$site_num++;
		for ($i = 0; $i < $ntaxa; ++$i){
			$Ns_per_sample[$i] = $Ns_per_sample[$i] + $N_for_site[$i];
		}
	}
	$i = 0;
	for ($i = 0; $i < $ntaxa; ++$i){
		$N_for_site[$i] = 0;
	}
	$num = 2;
	$N_num = 0;
	$i = 0;
}

for ($i = 0; $i < $ntaxa; ++$i){
	$prop_of_Ns[$i] = 0;
}

for ($i = 0; $i < $ntaxa; ++$i) {
	$prop_of_Ns[$i] = ($Ns_per_sample[$i] / $site_num);
	print "$header_stuff[($i+ 2)] = $prop_of_Ns[$i]\n";
}
```

Appendix 22 – Script to change LG of markers in a list, in a map check file. The map check file is a map file with chromosome and chromosome position information.

```
#! /usr/bin/perl -w
use strict;
use warnings;

# script changes the LG info in map_check file of markers (on given chromosome) to 0 if marker not # matching LG number given

open (INPUTFILE2, "<$ARGV[0]") or die "could not open map file from lepmap.\n";

my $chr_info = $ARGV[1] or die "enter chrom markers to change e.g Hmel221 or Hmel201.\n";
my $LG_info = $ARGV[2] or die "enter linkage group markers should be e.g 21 or 1";

my $line1; my @line_stuff1; my $marker_info1;

my $output_file1 = "$ARGV[0]_rm_markersmap";
open (OUTPUT, ">$output_file1");

my $zero = '0';

#read through file, and check marker info, if not matching, change to 0
while ($line1 = <INPUTFILE2>) {
	chomp $line1;
	@line_stuff1 = split'\t', $line1;
	$marker_info1 = substr $line_stuff1[0], 0, 7;
	  if ($marker_info1 eq $chr_info) {
		if ($line_stuff1[2] eq $LG_info) {
			print OUTPUT "$line1\n";
		}
		elsif ($line_stuff1[2] eq $zero) {
			print OUTPUT "$line1\n";
		}
		else {
			print OUTPUT "$line_stuff1[0]\t$line_stuff1[1]\t0\n";
		}
	  }
	  else {
		print OUTPUT "$line1\n";
	  }
}
```

Appendix 23 – Script to remove (set to 0) markers in a list from a map check file.

```
#! /usr/bin/perl -w
use strict;
use warnings;

# script changes the LG info in map_check file of markers (on given chromosome) to 0 if
marker not # matching LG number given

open (INPUTFILE1, "<$ARGV[0]") or die "could not open Chr : LG number list - must be
in order of Chr1-21.\n";
open (INPUTFILE2, "<$ARGV[1]") or die "could not open map check file from
lepmap.\n";

my $line; my @line_stuff;
my $LG_info; my $chr_info;
my $line1; my @line_stuff1; my $marker_info1;
my $first_mark = 0;
my $change = 0;

my $output_file1 = "$ARGV[1]_rm_markersmap";
open (OUTPUT, ">$output_file1");

my $zero = '0';

while ($line = <INPUTFILE1>) {
        chomp $line;
        @line_stuff = split'\t', $line;
        $chr_info = $line_stuff[0];
        $LG_info = $line_stuff[1];
          if ($first_mark != 0){
                  if ($line_stuff1[2] eq $LG_info) {
                  print OUTPUT "$line1\n";
              }
              elsif ($line_stuff1[2] eq $zero) {
                        print OUTPUT "$line1\n";
              }
              else {
                      print OUTPUT "$line_stuff1[0]\t$line_stuff1[1]\t0\n";
                          $change++;
                  }
          }
#read through file, and check marker info, if not matching, change to 0
          while ($line1 = <INPUTFILE2>) {
          chomp $line1;
          @line_stuff1 = split'\t', $line1;
          $marker_info1 = substr $line_stuff1[0], 0, 7;
                  if ($marker_info1 eq $chr_info) {
                          if ($line_stuff1[2] eq $LG_info) {
                                  print OUTPUT "$line1\n";
                          }
                          elsif ($line_stuff1[2] eq $zero) {
                                  print OUTPUT "$line1\n";
                          }
```

```
                        else {
                                print OUTPUT "$line_stuff1[0]\t$line_stuff1[1]\t0\n";
                                $change++;
                        }
                }
                else {
                        last;
                }
        }
        $first_mark++;
}
print "Changed Markers = $change\n";
```

Abbreviations

GC-MS – Gas chromatography coupled to mass spectrometry

GQ – Genotype quality

LG – Linkage group

PCA – Principal component analysis

PCR – Polymerase chain reaction

QTL – Quantitative trait loci

SNP – Single nucleotide polymorphism

TWISST – Topology weighting by iterative sampling of sub-trees

VCF – Variant call format

Bibliography

Altshuler D, Lander E, Ambrogio L (2010) A map of human genome variation from population scale sequencing. *Nature*, **476**, 1061–1073.

Andersson J, Borg-Karlson A-K, Vongvanich N, Wiklund C (2007) Male sex pheromone release and female mate choice in a butterfly. *The Journal of Experimental Biology*, **210**, 964–70.

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in speciation: Incipient speciation of sunflower ecotypes. *Evolution*, **67**, 2468–2482.

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

Arnegard ME, Kondrashov AS (2004) Sympatric speciation by sexual selection alone is unlikely. *Evolution*, **58**, 222–37.

Avery OT, Macleod CM, McCarty M (1944) Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. *Journal of Experimental Medicine*, **79**, 137–158.

Bao Y-Y, Qin X, Yu B *et al.* (2014) Genomic insights into the serine protease gene family and expression profile analysis in the planthopper, *Nilaparvata lugens*. *BMC Genomics*, **15**, 507.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, **439**, 719–23.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Review Genetics*, **12**, 767–780.

Bates D, Maechler M, Bolker BM, Walker S (2015) Fitting Linear Mixed-Effects Models using {lme4}. *Journal Of Statistical Software*, **67**, 1–48.

Baxter SW, Johnston SE, Jiggins CD (2008a) Butterfly speciation and the distribution of gene effect sizes fixed during adaptation. *Heredity*, **102**, 57–65.

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

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

Beldade P, Brakefield PM (2002) The genetics and evo-devo of butterfly wing patterns. *Nature Reviews Genetics*, **3**, 442–52.

Beldade P, Brakefield PM, Long AD (2002) Contribution of *Distal-less* to quantitative variation in butterfly eyespots. *Nature*, **415**, 315–318.

Van Belleghem SM, Rastas P, Papanicolaou A *et al.* (2016) Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Evolution and Ecology*, **1**, 1–32.

Berdan EL, Mazzoni CJ, Waurick I, Roehr JT, Mayer F (2015) A population genomic scan in *Chorthippus* grasshoppers unveils previously unknown phenotypic divergence. *Molecular Ecology*, 3918–3930.

Blum MJ (2008) Ecological and genetic associations across a *Heliconius* hybrid zone. *Journal of Evolutionary Biology*, **21**, 330–41.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 459–487.

Boughman JW, Rundle HD, Schluter D (2005) Parallel evolution of sexual isolation in sticklebacks. *Evolution*, **59**, 361–73.

Brakefield PM (1998) The evolution - development interface and advances with the eyespot patterns of *Bicyclus* butterflies. *Hereditas*, **80**, 265–272.

Brakefield PM, Gates J, Keys D *et al.* (1996) Development, plasticity and evolution of butterfly eyespot patterns. *Nature*, **384**, 236–242.

Brown KS (1976) An illustarted key to the silvaniform *Heliconius* (Lepidoptera: Nymphalidae) with descriptions of new subspecies. *Transactions of the American Entomological society*, **102**, 373–484.

Brown KS, Sheppard PM, Turner JRG (1974) Quaternary refugia in tropical America: evidence from race formation in *Heliconius* butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **187**, 369–378.

Brunetti CR, Selegue JE, Monteiro A *et al.* (2001) The generation and diversification of butterfly eyespot color patterns. *Current Biology*, **11**, 1578–1585.

Cannavò E, Khoueiry P, Garfield DA *et al.* (2016) Shadow enhancers are pervasive features of developmental regulatory networks. *Current Biology*, **26**, 38–51.

Chamberlain NL, Hill RI, Baxter SW, Jiggins CD, Kronforst MR (2011) Comparative population genetics of a mimicry locus among hybridizing *Heliconius* butterfly species. *Heredity*, **107**, 200–204.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science*, **327**, 302–5.

Chang CC, Chow CC, Tellier LC *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Cliften P (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5039–47.

Costanzo K, Monteiro A (2007) The use of chemical and visual cues in female choice in the butterfly *Bicyclus anynana*. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 845–51.

Costello MJ, May RM, Stork NE (2013) Can we name earth's species before they go extinct? *Science*, **339**, 413–416.

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

Coyne JA, Orr HA (2004) *Speciation*. Sunderland, MA: Sinauer Associates.

Cresko WA, Amores A, Wilson C *et al.* (2004) Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *PNAS*, **101**, 6050–6055.

Darwin C (1859) *On the Origin of Species by means of natural selection*. J. Murray, London.

Dasmahapatra KK, Walters JR, Briscoe AD *et al.* (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 2–6.

Davey JW, Chouteau M, Barker SL *et al.* (2016) Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3; Genes|Genomes|Genetics*, **6**, 695–708.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–8.

Dray S, Dufour AB (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22**, 1–20.

Dun G, Li X, Cao H, Zhou R, Li L (2007) Variations of *Melanocortin Receptor 1* (*Mc1R*) gene in three pig breeds. *Journal of Genetics and Genomics*, **34**, 777–782.

Eizirik E, Yuhki N, Johnson WE *et al.* (2003) Molecular genetics and evolution of melanism in the cat family. *Current Biology*, **13**, 448–453.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, **29**, 51–63.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends in Ecology and Evolution*, **26**, 298–306.

Eltringham H (1925) On the abdominal glands in *Heliconius* (Lepidoptera). *Transactions of the Entomological Society of London*, 269–275.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *PNAS*, **107**, 16196–16200.

Engler-Chaouat HS, Gilbert LE (2007) *De novo* synthesis vs. sequestration: negatively correlated metabolic traits and the evolution of host plant specialization in cyanogenic butterflies. *Journal of Chemical Ecology*, **33**, 25–42.

Estrada C, Jiggins CD (2002) Patterns of pollen feeding and habitat preference among *Heliconius* species. *Ecological Entomology*, **27**, 448–456.

Etter PD, Bassham S, Hohenlohe P a, Johnson E a, Cresko W a (2012) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics, Methods in Molecular Biology, vol. 772*, pp. 157–178.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, **35**, 124.

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

Ferguson LC, Maroja L, Jiggins CD (2011) Convergent, modular expression of ebony and tan in the mimetic wing patterns of *Heliconius* butterflies. *Development Genes and Evolution*, **221**, 297–308.

Fitzpatrick BM, Fordyce J a, Gavrilets S (2008) What, if anything, is sympatric speciation? *Journal of Evolutionary Biology*, **21**, 1452–9.

Frankel N, Wang S, Stern DL (2012) Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *PNAS*, **109**, 20975–9.

Gallant JR, Imhoff VE, Martin A *et al.* (2014a) Ancient homology underlies adaptive mimetic diversity across butterflies. *Nature Communications*, **5**, 1–10.

Gallant JR, Traeger LL, Volkening JD *et al.* (2014b) Genomic basis for the convergent evolution of electric organs. *Science*, **344**, 1522–1525.

Garrigan D, Kingan SB, Geneva AJ *et al.* (2012) Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Research*, **22**, 1499–511.

Gavrilets S (2004) *Fitness landscapes and the origin of species*. Princeton university press.

Geiger MF, McCrary JK, Schliewen UK (2010) Not a simple case - A first comprehensive phylogenetic hypothesis for the Midas cichlid complex in Nicaragua (Teleostei: Cichlidae: Amphilophus). *Molecular Phylogenetics and Evolution*, **56**, 1011–24.

Gilbert LE (2002) Adaptive novelty through introgression in *Heliconius* wing patterns: evidence for shared genetic "tool box" from synthetic hybrid zones and a theory of diversification. In: *Butterflies as model systems*, pp. 1–34.

Gompel N, Prud'homme B (2009) The causes of repeated genetic evolution. *Developmental Biology*, **332**, 36–47.

Gompert Z, Fordyce J a, Forister ML, Shapiro AM, Nice CC (2006) Homoploid hybrid speciation in an extreme habitat. *Science*, **314**, 1923–5.

Gould S (1990) *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton and Company, New York.

Gross JB, Borowsky R, Tabin CJ (2009) A novel role for *Mc1r* in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLoS Genetics*, **5**, e1000326.

Haffter P, Granato M, Brand M *et al.* (1996) The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development*, **123**, 1–36.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Syposium Series*, **41**, 95–98.

Hartlieb E, Anderson P (1999) Olfactory-Released Behaviours. In: *Insect olfaction* (ed Hanson BS), pp. 315–349. Springer, Berlin.

Hay-Roe MM, Nation J (2007) Spectrum of cyanide toxicity and allocation in *Heliconius erato* and *Passiflora* host plants. *Journal of chemical ecology*, **33**, 319–29.

Hedges SB (2002) The origin and evolution of model organisms. *Nature Reviews Genetics*, **3**, 838–849.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology*, **22**, 4606–18.

Hendry AP, Huber SK, De León LF, Herrel A, Podos J (2009) Disruptive selection in a bimodal population of Darwin's finches. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 753–759.

Hines HM, Counterman B a, Papa R *et al.* (2011) Wing patterning gene redefines the mimetic history of *Heliconius* butterflies. *PNAS,* **108**, 19666–19671.

Hines HM, Papa R, Ruiz M *et al.* (2012) Transcriptome analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing pattern variation. *BMC Genomics*, **13**, 288.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Hof AE van't, Campagne P, Rigden DJ *et al.* (2016) The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, **534**, 102–105.

Hong J-W, Hendrix DA, Levine MS (2008) Shadow enhancers as a source of evolutionary novelty. *Science*, **321**, 1314.

Hoyal Cuthill J, Charleston M (2012) Phylogenetic codivergence supports coevolution of mimetic *Heliconius* butterflies. *PLoS ONE*, **7**, e36464.

Huber B, Whibley A, Poul YL *et al.* (2015) Conservatism and novelty in the genetic architecture of adaptation in *Heliconius* butterflies. *Heredity*, **114**, 515–524.

Huxley J (1942) *Evolution: the modern synthesis*. Allen & Unwin, London.

Jang IH, Chosa N, Kim SH *et al.* (2006) A Spätzle-processing enzyme required for toll signaling activation in *Drosophila* innate immunity. *Developmental Cell*, **10**, 45–55.

Jiggins CD (2008) Ecological Speciation in Mimetic Butterflies. *BioScience*, **58**, 541–548.

Jiggins CD, Mavarez J, Beltrán M *et al.* (2005) A genetic linkage map of the mimetic butterfly *Heliconius melpomene*. *Genetics*, **171**, 557–70.

Jiggins CD, McMillan WO (1997) The genetic basis of an adaptive radiation: warning colour in two *Heliconius* species. *Proceedings of the Royal Society B: Biological Sciences*, **264**, 1167–1175.

Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature*, **411**, 302–305.

Jin YS, Anderson K V (1990) Dominant and recessive alleles of the *Drosophila* easter gene are point mutations at conserved sites in the serine protease catalytic domain. *Cell*, **60**, 873–881.

Jones RT, Salazar PA, Ffrench-Constant RH *et al.* (2012) Evolution of a mimicry supergene from a multilocus architecture. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 316–25.

Joron M, Frezal L, Jones RT *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Joron M, Jiggins CD, Papanicolaou A, Mcmillan WO (2006a) *Heliconius* wing patterns: an evo-devo model for understanding phenotypic diversity. *Heredity*, **97**, 157–67.

Joron M, Mallet JL (1998) Diversity in mimicry: paradox or paradigm? *Trends in Ecology and Evolution*, **13**, 461–6.

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

Kajitani R, Toshimoto K, Noguchi H *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.

Kapan DD (1998) Divergent natural selection and Müllerian mimicry in polymorphic *Heliconius cydno* (Lepidoptera: Nymphalidae). University of British Columbia.

Keller I, Wagner CE, Greuter L *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Kellogg DA, Doctor BP, Loebel JE, Nirenberg MW (1966) RNA codons and protein synthesis. IX. Synonym codon recognition by multiple species of valine-, alanine-, and methionine-sRNA. *PNAS*, **55**, 912–919.

Klein L, de Araújo AM (2010) Courtship behavior of *Heliconius erato phyllis* (Lepidoptera, Nymphalidae) towards virgin and mated females: conflict between attraction and repulsion signals? *Journal of Ethology*, **28**, 409–420.

Kozak KM, Wahlberg N, Neild AFE *et al.* (2015) Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. *Systematic Biology*, **64**, 505–524.

Kronforst MR, Kapan DD, Gilbert LE (2006a) Parallel genetic architecture of parallel adaptive radiations in mimetic *Heliconius* butterflies. *Genetics*, **174**, 535–549.

Kronforst MR, Papa R (2015) The functional basis of wing patterning in *Heliconius* butterflies: The molecules behind mimicry. *Genetics*, **200**, 1–19.

Kronforst MR, Salazar C, Linares M, Gilbert LE (2007) No genomic mosaicism in a putative hybrid butterfly species. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1255–1264.

Kronforst MR, Young LG, Kapan DD *et al.* (2006b) Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *PNAS*, **103**, 6575–6580.

Kryvokhyzha D (2014) Whole genome resequencing of *Heliconius* butterflies revolutionizes our view of the level of admixture between species. *Master's Thesis in Evolutionary Biology, Uppsala University & Harvard University*, 30.

Kuhn TS (1970) *The Structure of Scientific Revolutions*. University of Chicago Press.

Lamichhaney S, Berglund J, Almén MS *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

Lander ES, Linton LM, Birren B *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lento W, Congdon K, Voermans C *et al.* (2012) *Wnt / Wingless* Signaling in *Drosophila*. *Cold Spring Harb Perspect Biol*, **4**, a007930.

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

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liénard M a, Wang H-L, Lassance J-M, Löfstedt C (2014) Sex pheromone biosynthetic pathways are conserved between moths and the butterfly *Bicyclus anynana*. *Nature Communications*, **5**, 3957.

Linares M (1996) The genetics of the mimetic coloration in the butterfly *Heliconius cydno* weymeri. *Heredity*, **87**, 142–149.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *PNAS*, **113**, 5970–5975.

Lofstedt C (1993) Moth pheromone genetics and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **340**, 167–177.

Mallet J (1986) Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity*, **56**, 191–202.

Mallet J (1989) The genetics of warning clour in Peruvian hybrid zones of *Heliconius erato* and *H. melpomene*. *Proceedings of the Royal Society B: Biological Sciences*, **236**, 163–185.

Mallet J (2006) What does *Drosophila* genetics tell us about speciation? *Trends in Ecology and Evolution*, **21**, 386–393.

Mallet J, Barton NH (1989a) Strong natural selection in a warning-color hybrid zone. *Evolution*, **43**, 421–431.

Mallet J, Barton N (1989b) Inference from clines stabilized by frequency-dependent selection. *Genetics*, **122**, 967–76.

Mallet J, Joron M (1999) Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. *Annual Review of Ecology and Systematics*, **30**, 201–233.

Mallet J, Meyer A, Nosil P, Feder JL (2009) Space, sympatry and speciation. *Journal of Evolutionary Biology*, **22**, 2332–41.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

Martin CH, Cutler JS, Friel JP *et al.* (2015a) Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, **69**, 1406–1422.

Martin SH, Davey JW, Jiggins CD (2014a) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology Evolution*, **32**, 244–257.

Martin SH, Eriksson A, Kozak KM, Manica A, Jiggins CD (2015b) Speciation in *Heliconius* butterflies: minimal contact followed by millions of generations of hybridisation. *BioRxiv*, 1–24.

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

Martin A, Papa R, Nadeau NJ *et al.* (2012) Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. *PNAS*, **109**, 12632–12637.

Martin A, Reed RD (2010) *Wingless* and *aristaless2* define a developmental ground plan for moth and butterfly wing pattern evolution. *Molecular Biology and Evolution*, **27**, 2864–2878.

Mattersdorfer K, Koblmüller S, Sefc KM (2012) AFLP genome scans suggest divergent selection on colour patterning in allopatric colour morphs of a cichlid fish. *Molecular Ecology*, **21**, 3531–3544.

Matthaei JH, Jones OW, Martin RG, Nirenberg MW (1962) Characteristics and composition of RNA coding units. *PNAS*, **48**, 666–77.

Mavárez J, Salazar CA, Bermingham E *et al.* (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, **441**, 868–71.

Mayr E (1963) *Animal Species and Evolution*. Cambridge, MA: Belknap.

McGregor AP, Orgogozo V, Delon I *et al.* (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, **448**, 587–590.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–303.

McMillan WO, Jiggins CD, Mallet J (1997) What initiates speciation in passion-vine butterflies? *PNAS*, **94**, 8628–8633.

McMillan WOO, Monteiro A, Kapan DD (2002) Development and evolution on the wing. *Trends in Ecology and Evolution*, **17**, 125–133.

Merrill RM, Chia A, Nadeau NJ (2014) Divergent warning patterns contribute to assortative mating between incipient *Heliconius* species. *Ecology and Evolution*, **4**, 911–917.

Merrill RM, Dasmahapatra KK, Davey JW *et al.* (2015) The diversification of *Heliconius* butterflies: What have we learned in 150 years? *Journal of Evolutionary Biology*, **28**, 1417–1438.

Merrill RM, Van Schooten B, Scott JA, Jiggins CD (2011) Pervasive genetic associations between traits causing reproductive isolation in *Heliconius* butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 511–8.

Merrill RM, Wallbank RWR, Bull V *et al.* (2012) Disruptive ecological selection on a mating cue. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 4907–13.

Meyer A, Kocher TD, Basasibwaki P, Wilson AC (1990) Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature*, **347**, 550–553.

Monteiro A, Brakefield PM, French V (1997) The genetics and development of an eyespot pattern in the butterfly *Bicyclus anynana*: response to selection for eyespot shape. *Genetics*, **146**, 287–294.

Monteiro A, Chen B, Ramos DM *et al.* (2013) *Distal-less* regulates eyespot patterns and melanization in *Bicyclus* butterflies. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, **320**, 321–331.

Monteiro A, Podlaha O (2009) Wings, horns, and butterfly eyespots: How do complex traits evolve? *PLoS Biology*, **7**, 0209–0216.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? *PLoS Biology*, **9**, 1–8.

Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, **155**, 335–350.

Mundy NI (2005) A window on the genetics of evolution: *Mc1r* and plumage colouration in birds. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 1633–1640.

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

Nadeau NJ, Pardo-Diaz C, Whibley A *et al.* (2016) The gene *cortex* controls mimicry and crypsis in butterflies and moths. *Nature*, **534**, 106–110.

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

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

Naisbit RE, Jiggins CD, Mallet J (2001) Disruptive sexual selection against hybrids contributes to speciation between *Heliconius cydno* and *Heliconius melpomene*. *Proceedings of the Royal Society B: Biological Sciences*, **268**, 1849–54.

Nijhout HF (1990) A comprehensive model for colour pattern formation in butterflies. *Proceedings of the Royal Society B: Biological Sciences*, **239**, 81–113.

Nijhout H (1994) Symmetry systems and compartments in Lepidopteran wings: the evolution of a patterning mechanism. *Development*, **233**, 225–233.

Nirenberg MW, Matthaei JHH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *PNAS*, **47**, 1588–602.

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

Orr HA (2005a) The Probability of Parallel Evolution. *Evolution*, **59**, 216–220.

Orr HA (2005b) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, **6**, 119–127.

Osanai-Futahashi M, Tatematsu K, Futahashi R *et al.* (2016) Positional cloning of a *Bombyx* pink-eyed white egg locus reveals the major role of *cardinal* in ommochrome synthesis. *Heredity*, **116**, 135–145.

Owen R (1848) *On the Archetype and Homologies of the Vertebrate Skeleton*. Van Voorst.

Papa R, Kapan DD, Counterman BA *et al.* (2013) Multi-Allelic Major Effect Genes Interact with Minor Effect QTLs to Control Adaptive Color Pattern Variation in *Heliconius erato*. *PLoS ONE*, **8**, e57033.

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

Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445–464.

Parker J, Tsagkogeorga G, Cotton J a *et al.* (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**, 1–9.

Penz CM, Krenn HW (2000) Behavioral adaptations to pollen-feeding in *Heliconius* butterflies (Nymphalidae , Heliconiinae): an experiment using lantana flowers. *Journal of Insect Behavior*, **13**, 865–880.

Perez SI, Bernal V, Gonzalez PN (2006) Differences between sliding semi-landmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. *Journal of Anatomy*, **208**, 769–84.

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. *Genomics*, **93**, 105–111.

Pigliucci M (2007) Do we need an extended evolutionary synthesis? *Evolution*, **61**, 2743–2749.

Pohlert T (2014) The pairwise multiple comparison of mean ranks package (PMCMR).

Powell JR (1997) *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press, Oxford.

Protas M, Conrad M, Gross JB, Tabin C, Borowsky R (2007) Regressive evolution in the mexican cave tetra, Astyanax mexicanus. *Current Biology*, **17**, 452–454.

Protas ME, Hersey C, Kochanek D *et al.* (2006) Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature genetics*, **38**, 107–11.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, **16**, 359–371.

Rastas P, Calboli FCF, Guo B, Shikano T, Merilä J (2016) construction of ultradense linkage maps with Lep-MAP2: stickleback f2 recombinant crosses as an example. *Genome Biology and Evolution*, **8**, 78–93.

Reed RD, McMillan WO, Nagy LM (2008) Gene expression underlying adaptive variation in *Heliconius* wing patterns: non-modular regulation of overlapping *cinnabar* and *vermilion* prepatterns. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 37–45.

Reed RD, Nagy LM (2005) Evolutionary redeployment of a biosynthetic module: expression of eye pigment genes *vermilion, cinnabar,* and *white* in butterfly wing development. *Evolution and Development*, **7**, 301–311.

Reed RD, Papa R, Martin A *et al.* (2011) Optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, **333**, 1137–1141.

Reichhart JM, Gubb D, Leclerc V (2011) *The Drosophila serpins: Multiple functions in immunity and morphogenesis*. *Methods in Enzymology*, **499**, 205-225

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1–8.

Rodrigues Macedo M, Das Graças M, Freire M (2011) Insect digestive enzymes as a target for pest control. *Isj*, **8**, 190–198.

Rogers SM, Bernatchez L (2004) FAST-TRACK: Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **14**, 351–361.

Rohlf JF (2013a) tpsDig, digitize landmarks and outlines.

Rohlf JF (2013b) tpsUtil, file utility program.

Römpler H, Rohland N, Lalueza-Fox C *et al.* (2006) Nuclear gene indicates coat-color polymorphism in mammoths. *Science*, **313**, 62.

Ross J, Jiang H, Kanost MR, Wang Y (2003) Serine proteases and their homologs in the *Drosophila* melanogaster genome: an initial analysis of sequence conservation and phylogenetic relationships. *Gene*, **304**, 117–131.

Rosser N, Dasmahapatra KK, Mallet J (2014) Stable *Heliconius* butterfly hybrid zones are correlated with a local rainfall peak at the edge of the Amazon basin. *Evolution*, 1–39.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. *Molecular Biology and Evolution*, **30**, 1574–87.

Rutowski RL (1980) Male scent-producing structures in *Colias* butterflies - Function, localization, and adaptive features. *Journal of Chemical Ecology*, **6**, 13–26.

Saiki R, Scharf S, Faloona F *et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–1354.

Salazar C a, Jiggins CD, Arias CF *et al.* (2005) Hybrid incompatibility is consistent with a hybrid origin of *Heliconius heurippa* (Hewitson) from its close relatives, *Heliconius cydno* (Doubleday) and *Heliconius melpomene* (Linnaeus). *Journal of Evolutionary Biology*, **18**, 247–256.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**, 441–448.

Savolainen V, Anstett M-C, Lexer C *et al.* (2006) Sympatric speciation in palms on an oceanic island. *Nature*, **441**, 210–213.

Schliewen UK, Klee B (2004) Reticulate sympatric speciation in Cameroonian crater lake cichlids. *Frontiers in Zoology*, **1**, 5.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science*, **323**, 737–741.

Schulz S, Estrada C, Yildizhan S, Boppré M, Gilbert LE (2008) An antiaphrodisiac in *Heliconius melpomene* butterflies. *Journal of Chemical Ecology*, **34**, 82–93.

Servedio MR, Doorn GS Van, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: "magic" but not rare? *Trends in Ecology and Evolution*, **26**, 389–397.

Shapiro MD, Marks ME, Peichel CL *et al.* (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**, 717–723.

Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. *Nature reviews. Genetics*, **5**, 335–344.

Sheppard PM (1963) Some genetic studies of Müllerian mimics in butterflies of the genus *Heliconius*. *Zoologica*, **48**, 145–154.

Sheppard PM, Turner JRG, Brown KS *et al.* (1985) Genetics and the evolution of Müllerian mimicry in *Heliconius* butterflies. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **308**, 433–610.

Sherratt TN (2006) Spatial mosaic formation through frequency-dependent selection in Müllerian mimicry complexes. *Journal of Theoretical Biology*, **240**, 165–74.

Shoup JR (1966) The development of pigment granules in the eyes of wild type and mutant *Drosophila melanogaster*. *Journal of Cell Biology*, **29**, 223–249.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM (2009) ABySS : A parallel assembler for short read sequence data. *Genome Research*, **19**, 1117–1123.

Smith HO, Welcox, K W (1970) A restriction enzyme from *Hemophilus influenzae*: I. purification and general properties. *Journal of Molecular Biology*, **51**, 379–391.

Song Y, Endepols S, Klemann N *et al.* (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, **21**, 1296–1301.

Soria-Carrasco V, Gompert Z, Comeault A a *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–42.

Stark WS, Srygley RB, Greenberg RM (1981) Analysis of a compound eye mosaic of outer rhabdomeres absent marked with cardinal. *Drosophila Information Servic*, **56**, 132–133.

Stavenga DG (2002) Colour in the eyes of insects. *Journal of comparative physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology*, **188**, 337–348.

Stern DL (2013) The genetic causes of convergent evolution. *Nature Reviews Genetics*, **14**, 751–64.

Stern DL, Orgogozo V (2009) Is genetic evolution predictable? *Science*, **323**, 746–751.

Stucky BJ (2012) Seqtrace: A graphical tool for rapidly processing DNA sequencing chromatograms. *Journal of Biomolecular Techniques*, **23**, 90–93.

Summers K, Symula R, Clough M, Cronin T (1999) Visual mate choice in poison frogs. *Proceedings of the Royal Society B: Biological Sciences*, **266**, 2141–2145.

Supple MA, Hines HM, Dasmahapatra KK *et al.* (2013) Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome research*, **23**, 1248–1257.

Symonds MRE, Elgar MA (2008) The evolution of pheromone diversity. *Trends in Ecology and Evolution*, **23**, 220–228.

Taylor M V. (2002) *Drosophila* development: Novel signal dispatch elicits visceral response. *Current Biology*, **12**, 102–104.

Tearle R (1991) Tissue specific effects of ommochrome pathway mutations in *Drosophila melanogaster*. *Genet Res*, **57**, 257–266.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, **22**, 4673–80.

Tishkoff S a, Reed F a, Ranciaro A *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, **39**, 31–40.

Turner JRG (1977) Butterfly mimicry: the genetical evolution of an adaptation. *Evolutionary Biology*, **11**, 163–206.

Turner JRG (1981) Adaptation and evolution in *Heliconius*: a defense of NeoDarwinism. *Annual Review of Ecology and Systematics*, **12**, 99–121.

Turner JRG, Crane J (1962) The genetics of some polymorphic forms of the butterflies *Heliconius melpomene* (Linneaus) and *H. erato* (Linneaus), I: major genes. *Zoologica*, **1962**, 141–152.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–89.

Valverde P, Healy E, Jackson I, Rees JL, Thody a J (1995) Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature genetics*, **11**, 328–330.

Vanjari S, Mann F, Merrill RM, Schulz S, Jiggins CD (2015) Male sex pheromone components in the butterfly *Heliconius melpomene*. *BioRxiv*, 33506.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 451–460.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334–4345.

Wallbank RWR, Baxter SW, Pardo-Díaz C *et al.* (2016) Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biology*, 1–16.

Watson JD, Crick FH (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.

Weinreich DM (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, **312**, 111–114.

Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics*, **13**, 59–69.

Wray GA. (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, **8**, 206–216.

Wu C (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu CI, Hollocher H, Begun DJ *et al.* (1995) Sexual isolation in *Drosophila* melanogaster: a possible case of incipient speciation. *PNAS*, **92**, 2519–2523.

Ye J, Coulouris G, Zaretskaya I *et al.* (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.

Zhang J (2006) Parallel adaptive origins of digestive RNAses in Asian and African leaf monkeys. *Nat Genet*, **38**, 819–823.

Zhang L, Reed RD (2016) Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns. *Nature Communications*, **7**, 1–7.

Zou Z, Lopez DL, Kanost MR, Evans JD, Jiang H (2006) Comparative analysis of serine protease-related genes in the honey bee genome: possible involvement in embryonic development and innate immunity. *Insect Molecular Biology*, **15**, 603–614.