

# *Hox3* Duplication and Divergence in the Lepidoptera



Luca Livraghi

PhD Thesis



# *Hox3* Duplication and Divergence in the Lepidoptera

Luca Livraghi

Oxford Brookes University

Thesis submitted in partial fulfilment of the requirements for  
the award of Doctor of Philosophy

First submitted August 2017

# Acknowledgements

This thesis was written with the hope that it will inspire (or at least somewhat entertain) future researchers. There are many people who were instrumental to many aspects of this work and for their help, support and assistance I'd like to thank the following people.

I am thankful to my supervisors Dr Casper Breuker and Dr Melanie Gibbs for their guidance and support, for fostering a stimulating scientific environment, and for their continuous help in writing this document. I also thank my co-supervisor, Professor Peter Holland for stimulating conversation and direction.

I also thank my collaborators; Dr Ben Longdon, for a collaboration on the first description of rhabdovirus found in Speckled Wood populations. Dr Leonardo Dapporto, Dr Raluca Voda and Dr Roger Vila, with whom we described the biogeography of *Pararge aegeria*, and for their hospitality in Barcelona. Many thanks to Dr Arnaud Martin for the invaluable help in establishing the CRISPR/Cas9 technique.

I would also like to thank past and current members of the Breuker Research group. Dr Jean-Michel Carter, for his help with *in situ* hybridisations and general guidance during the start of my PhD. Nora Braak, for images of butterfly embryos and helpful discussions. I wish her all the best with her PhD and future endeavours. Luke Evans for his invaluable contribution to the mating experiments, and Mimi Phung, for assistance with the CRISPR. Many thanks also to past and current members of the McGregor lab, for always being there for helpful comments and discussion and, most importantly, for the copious amount of alcohol consumption. In no particular order, thank you Daniel Leite, Dr Anna Schönauer, Dr Isabel Almudi, Christine Ashton, Michaela Holzem, Dr Pedro Gaspar, Alexandra Buffry, Dr Kentaro Tanaka, Dr Cláudia Mendes, Dr Sebastian Kittleman and Professor Alistair McGregor.

I would like to extend a special gratitude to Dr Daniela Nunes and Dr Saad Arif. To Daniela for her passionate discussions and help, and to Saad for invaluable help in establishing gene editing tools. Most importantly, to both of them for having shared many laughs (and cigarettes).

Thank you to all my friends who have kept me sane (or at least tried) throughout this whole process. Through fear of potential consequences, I name them each below. Dr Magdalena Svensson, Andrew Walmsley, Lucy Radford, Helen Bersacola, Tim Robbins, Gareth Lloyd, Albie Henry, Giacomo Montereale Gavazzi, Tom Cummings, Milena Pennec, Gaia Joloidovsky, Naomi Hendy, Lydia Luncz, Nick James, Clare Belden, Danny Berg and Katie Reinhardt. And all the friends from The Island (there are too many to name here).

Finally, thank you to my family. Asli Tatliadim, for an incredible journey. Annarella Rando, for being the best Italian mum (no bias here). Marco Livraghi, to whom I owe my curiosity, skepticism, and so much more. To Nadine Livraghi, whose travels I am so jealous of and to Pietro Maggiora, a scientist in the making.

# Publications

Longdon B., Day J., Schulz N., Leftwich P.T., de Jong M.A., Breuker C.J., Gibbs M., Obbard D.J., Wilfert L., Smith S.C.L., McGonigle J.E, Houslay T.M., Wright L.I., **Livraghi L.**, Evans L.C, Friend L.A., Chapman T., Vontas J., Kambouraki N. and Jiggins F.M. (2017). Vertically transmitted rhabdoviruses are found across three insect families and have dynamic interactions with their hosts. *Proceedings of the Royal Society B: Biological Sciences*, **284**(1847):20162381.

Mazo-Vargas A., Concha C., **Livraghi L.**, Massardo D., Wallbank R., Zhang L., Papador J., Martinez-Najera D., Jiggins C., Kronforst M., Breuker C.J., Reed R., Patel N., McMillan W., Martin A. (2017). Macro-evolutionary shifts of *WntA* function potentiate butterfly wing pattern diversity. *Proceedings of the National Academy of Sciences*, **114**(40):10701–10706.

**Livraghi L.**, Vodă R., Evans L.C, Gibbs M., Dincă V., Holland P., Shreeve T.G., Vila R., Dapporto L. and Breuker C.J. (Submitted to *Journal of Biogeography*). Disentangling patterns of historical and current gene flow in the butterfly *Pararge aegeria*.

**Livraghi L.**, Martin A., Gibbs M., Braak N., Arif S., and Breuker C.J. (Submitted to *Advances in Insect Physiology*). CRISPR/Cas9 as the key to unlocking the secrets of butterfly wing pattern development and its evolution.



# Abstract

Using the Speckled Wood Butterfly *Pararge aegeria* as the model species, this thesis presents the possible evolutionary significance of a set of duplications found in the Hox cluster of the Lepidoptera, called the Special Homeobox genes. An annotation of this duplicated cluster across a wide number of Lepidoptera was performed in order to assess patterns of duplication and loss across the order. The sequences recovered revealed a large amount of variation associated with the duplicate genes, indicating these are evolving very rapidly in different lineages. Patterns of sequence variation were examined to ascertain whether the observed variation was maintained due to selection at three separate levels of divergence: within the Ditrysia, within the more recently diverged *Heliconius* genus, and at the intraspecific level by quantifying nucleotide polymorphism within *Pararge aegeria*. Selective pressures were found to be operating between paralogous and orthologous genes, suggesting these have evolved, in part, under positive selection. The potential function of the duplicates was examined by means of CRISPR/Cas9 genome editing, but revealed inconclusive results. Genome editing, however, was shown to be largely applicable to *P. aegeria*, and resulted in consistent mutations associated with wing patterning genes. The potential significance of the duplications for Lepidopeteran biology are discussed, as well as future applications for genome editing techniques in *P. aegeria*.





Dedicated to Giancarlo Livraghi



# Table of Contents

<b>Acknowledgements</b> .....	4
<b>Publications</b> .....	6
<b>Abstract</b> .....	8
<b>Table of Contents</b> .....	12
<b>List of Figures</b> .....	16
<b>List of Tables</b> .....	18
<b>Chapter I - Introduction</b> .....	20
I.1. Introduction.....	21
I.1.a. Background.....	21
I.1.b. Deep homology, pleiotropy and <i>cis</i> -regulatory evolution.....	23
I.1.c. Gene Duplication and divergence.....	27
I.1.d. Hox3 duplications and the evolution of extraembryonic tissue in insects.....	33
I.1.e The Speckled Wood Butterfly, <i>Pararge aegeria</i> .....	44
I.1.f Development of gene editing tools in <i>Pararge aegeria</i> .....	46
I. 2. Aims.....	49
I.2.a. Research objective 1 - Interspecific divergence and selection on Hox3 locus across the Lepidoptera.....	50
I.2.b. Research objective 2 – Geographical distribution of polymorphisms associated with the Hox3 locus in <i>Pararge aegeria</i> .....	51
I.2.c. Research objective 3 – Development of CRISPR/Cas9 technology in <i>Pararge aegeria</i> .....	50
<b>Chapter II – Hox3 polymorphism and signatures of selection in the Lepidoptera</b> .....	51
II.1. Introduction.....	53
II.2. Materials and Methods.....	57
II.2.a Annotation of <i>Shx</i> genes in Lepidoptera and comparative analysis between species.....	57
II.2.b DNA extraction, amplification, sequencing, and alignment in <i>Pararge aegeria</i> .....	57
II.2.c Population genetic data from <i>Heliconius erato</i> and <i>Heliconius himera</i> .....	58
II.2.d Phylogenetic analyses.....	58
II.2.c Selection analyses.....	59

II.3. Results.....	61
II.3.a Molecular Evolution of the <i>Hox3</i> Locus in Lepidoptera.....	61
II.3.b Molecular Evolution of the <i>Hox3</i> Locus within species.....	64
II.4. Discussion.....	67
<b>Chapter III - <i>Pararge aegeria</i> displays patterns of geographic variation in the novel <i>Shx</i> genes distinct from its biogeography as inferred by <i>COI</i>.....</b>	<b>74</b>
III.1. Introduction.....	76
III.2. Material and Methods.....	80
III.2.a Study species.....	80
III.2.b DNA extraction, amplification, sequencing, and alignment.....	81
III.2.c Phylogenetic analyses.....	82
III.2.d Haplotype Networks and Genetic Landscapes.....	82
III.2.e Pre-zygotic reproductive barriers: Courtship behaviour in Sardinian and Corsican <i>Pararge aegeria</i> .....	83
III.2.f Reproductive barriers.....	85
III.2.g Backcrosses.....	85
III.2.h <i>Wolbachia</i> .....	86
III.2.i Statistical analyses Mating Experiments.....	86
III.3. Results.....	87
III.3.a <i>COI</i> variation reveals the presence of two distinct lineages.....	87
III.3.b Nuclear genes versus <i>COI</i> lineages.....	87
III.3.c Patterns of polymorphisms associated with the <i>Shx</i> genes.....	91
III.3.d Pre-zygotic reproductive barriers: Courtship behaviour.....	93
III.3.e Reproductive barriers.....	94
III.3.f Post-zygotic reproductive barriers.....	95
III.4. Discussion.....	95
<b>Chapter IV - Development of CRISPR/Cas9 technology in <i>Pararge aegeria</i>.....</b>	<b>101</b>
IV.1. Introduction.....	103
IV.2. Methods.....	109
IV.2.a Animal Husbandry.....	109
IV.2.b Cas9-mediated genome editing.....	110
IV.2.c Imaging.....	111

IV.2.d Geometric Morphometrics.....	111
IV.2.e <i>In Situ</i> Hybridisation of <i>Shx</i> genes.....	113
IV.3. Results.....	114
IV.3.a <i>yellow</i> Knock-outs in <i>Pararge aegeria</i> wings.....	115
IV.3.b <i>WntA</i> Knock-outs in <i>Pararge aegeria</i> wings.....	118
IV.3.c <i>WntA</i> and <i>yellow</i> effects on shape and size of <i>Pararge aegeria</i> wings – Geometric Morphometrics.....	121
IV.3.d Knock-out of <i>ShxA</i> and <i>ShxC</i> in <i>Pararge aegeria</i> .....	123
IV.3.e <i>ShxC</i> expression in embryonic tissue.....	126
IV.4. Discussion.....	127
IV.4.a Yellow has a conserved pigmentation role in <i>Pararge aegeria</i> .....	128
IV.4.b Evolutionary ecology of <i>Pararge aegeria</i> wing melanisation.....	129
IV.4.c <i>WntA</i> has a conserved role in wing patterning in butterflies.....	130
IV.4.d CRISPR/Cas9 injections may have an effect on overall wing size and shape...133	
IV.4.e <i>Shx</i> genes are likely pleiotropic.....	135
IV.4.f Future Prospects.....	137
IV.4.g Conclusions.....	138
<b>Chapter V – Conclusions</b> .....	139
V.1. <i>Pararge aegeria</i> , an emerging model species.....	140
V.2. Duplication and Divergence at the <i>Hox3</i> Locus and the Evolutionary Success of the Lepidoptera.....	142
V.3. Future Directions.....	146
<b>Bibliography</b> .....	148
<b>Appendices</b> .....	176
Appendix II.....	177
Appendix II - Figure 1. Tree topology based on Homeodomain alignments used in selection analyses (BS-REL).....	178
Appendix II – Figure 2. Putative location of positively selected sites as identified by MEME on homeodomain alignments spanning all Lepidoptera.....	179
Appendix II – Figure 3. Putative location of positively selected sites as identified by MEME on homeodomain alignments spanning <i>Heliconius</i> .....	180
Appendix II -Table 1. Homeodomain replacement polymorphisms in <i>Pararge aegeria</i> .....	181

Appendix II - Table 2. Homeodomain replacement polymorphisms in <i>Heliconius erato</i> .....	180
Appendix II - Table 3. Annotation of the <i>Hox3</i> locus in the Leiodopteran species analysed.....	182
Appendix II – Table 4. Summary of MEME analysis on Lepidopteran <i>Shx</i> genes.....	183
Appendix II – Table 5. Summary of sites under selection in <i>Heliconius</i> using all methods.....	184
Appendix III.....	187
Appendix III – Note 1. Primers and cycling conditions for nuclear genes.....	187
Appendix III Table 1 – <i>Wolbachia</i> screen.....	188
Appendix IV.....	189
Appendix IV - Figure 1. Ventral side of the <i>Pararge aegeria</i> wing surface highlighting landmarks used for the geometric morphometrics.....	190
Appendix IV - Figure 2. <i>Pararge aegeria</i> wild-type and <i>WntA</i> mKO mutants.....	190
Appendix IV - Figure 3. <i>Pararge aegeria</i> wild-type and <i>yellow</i> mKO mutants.....	191
Appendix IV – Figure 4. Phylogenetic analysis of <i>WntA</i> in <i>Pararge aegeria</i> .....	192
Appendix IV – Figure 5. Summary of significant effects reported on wing shape.....	193
Appendix IV – Additional file 1. CRISPR/Cas9 Protocol for <i>Pararge aegeria</i> .....	194
Appendix IV - Table 1. sgRNA target sequences and genotyping primers.....	197
Appendix IV - Table 2. Asymmetry of overall wing size and shape for internal and external landmarks in forewings and hindwings.....	198

# List of Figures

Figure I - 1 Possible fates of duplicate genes.....	30
Figure I - 2 <i>Hox</i> cluster expansion in arthropods.....	36
Figure I - 3 Duplications and functional divergence in <i>Hox3</i> genes across the arthropod phylogeny.....	38
Figure I - 4 Localisation of <i>Shx</i> transcripts in <i>Pararge aegeria</i> ovarioles (A-E), 10h embryos (F-J) and 12h embryos (K-O).....	42
Figure I - 5 Early embryonic development in <i>Pararge aegeria</i> .....	43
Figure I - 6 The nymphalid groundplan.....	48
Figure II - 1 The <i>Hox3</i> Locus in Lepidoptera.....	56
Figure III - 1 Distribution of <i>COI</i> haplotypes across Europe.....	89
Figure III - 2 Genetic Heterogeneity of <i>COI</i> in <i>Pararge aegeria</i> .....	90
Figure III - 3 Geographic clustering of nuclear genes in <i>Pararge aegeria</i> .....	91
Figure III - 4 Differential patterns of polymorphisms associated with the <i>Hox3</i> locus in <i>Pararge aegeria</i> .....	92
Figure III - 5 Possible recolonisation scenario for <i>Pararge aegeria</i> .....	97
Figure IV- 1 Phylogenetic analysis of the <i>yellow</i> family genes in <i>Pararge aegeria</i> .....	116
Figure IV- 2 Knock-out of the <i>yellow</i> locus in <i>Pararge aegeria</i> .....	117
Figure IV- 3 <i>WntA</i> Knock-out in <i>Pararge aegeria</i> .....	120
Figure IV- 4 Effects of <i>WntA</i> mKO on marginal elements of <i>Pararge aegeria</i> .....	121
Figure IV - 5 Effects of CRISPR/Cas9 injections on overall wing size in <i>Pararge aegeria</i> ....	122
Figure IV- 6 CRISPR/Cas9 injections for <i>Shx</i> in <i>Pararge aegeria</i> .....	125
Figure IV- 7 Targeted locations of the <i>ShxA</i> and <i>ShxC</i> loci in <i>Pararge aegeria</i> .....	126
Figure IV - 8 <i>ShxC</i> expression in <i>Pararge aegeria</i> embryos.....	127





# List of Tables

Table II - 1 Positive selection following gene duplication at the <i>Hox3</i> locus in <i>Ditrysia</i> .....	62
Table II - 2 Summary of Selection acting on the <i>Hox3</i> genes in <i>Heliconius</i> butterflies.....	64
Table II - 3 Sites across the <i>Heliconius</i> homeodomains under positive diversifying selection as identified by MEME.....	64
Table II - 4 Summary statistics and selection tests at each of the <i>Hox3</i> paralogs in <i>Pararge aegeria</i> .....	66
Table II - 5 Summary statistics and selection tests at each of the <i>Hox3</i> paralogs in <i>Heliconius erato</i> .....	67
Table III - 1 Collection details of the 22 females caught in the field and whose offspring were used in the laboratory crosses.....	81
Table III - 2 Crosses used in the mating experiments.....	84

“My own suspicion is that the world is not only queerer than we suppose,  
but queerer than we can suppose.”

J. B. S. Haldane, *Possible Worlds* - 1928

# Chapter I

## Introduction

## I.1. Introduction

### I.1.a Background

How changes in genetic information are interpreted by developmental programs to give rise to organismal diversity is the central problem in the field of Evolutionary Developmental Biology (Evo-Devo). Developmental networks translate genotypes into phenotypes, through their interaction with the environment, resulting in morphological and behavioural variation that can subsequently be shaped by natural selection (Gould, 1985; Klingenberg, 1998; Cheate Jarvela and Pick, 2016). Identifying the properties of development that directly underpin evolutionary change is essential to gain an understanding of how morphological diversity arises. In order to understand how developmental networks interact with each other and evolve, it is essential to disentangle the relative contribution of the individual components of such networks; namely, genes. Early experiments pioneered by Lewis, Gehring and others in *Drosophila melanogaster* (Lewis, 1978; Gehring, 1985; Gehring and Hiromi, 1986; Lewis, 1994; Lewis, 1998), were among the first to elucidate some of the key genes involved in the instruction of the genetic pathways leading to morphological differentiation. They were interested in mutations that resulted in different segments being transformed into a fully formed adjacent body part, such as halteres being transformed into wings, or antennae into legs in flies (Bateson, 1894). These homeotic transformations eventually led to the isolation of what came to be known as *Hox* genes, and comparative studies during the 1980s and 1990s led to the incredible discovery that *Hox* genes were shared amongst almost all animals (McGinnis *et al.*, 1984; Müller *et al.*, 1984; Levine *et al.*, 1984; Carrasco *et al.*, 1984; Holland and Hogan, 1988; McGinnis and Krumlauf, 1992; Holland, 2013). The realisation that these genes played a conserved role in embryonic segment identity specification amongst many different types of animals, and that their chromosomal clustering and expression (co-linearity) was also conserved (reviewed in Gaunt, 2015), re-shaped the way many biologists thought about development as well as the evolution of morphological complexity, and opened up the possibility of integrating an evolutionary perspective into the field.

The study of morphological variation and evolution, albeit being central themes to modern Biology, were based on historically disjointed fields (Bolker, 2008). Early naturalists primarily focused on the contribution of biological variation in an effort to classify the observed patterns of variation into taxonomic units. By the early 1900s, Darwin's ideas on natural selection became united with Mendel's principles of inheritance, and genes came to be recognised as the raw material for modification by descent (Darwin, 1859; Mendel *et al.*, 1866). During the early 20<sup>th</sup> Century, population geneticists formalised Mendel's and Darwin's theories and showed that genetic variation was the substrate for evolutionary innovation: alleles and genetic variants segregating through populations are shaped by selection, gene flow, demography and drift and ultimately result in morphological change and speciation. Thus, what became known as the Modern Synthesis of Evolution was born.

However, their theories largely considered genes as generic and abstract entities and mainly regarded their influence in terms of their effect on natural variation (*e.g.* additive, dominant or epistatic effects; the field of quantitative genetics) (Falconer and Mackay, 1996; Stern and Orgogozo, 2009). Although the exact molecular basis was poorly understood, it was recognised that the effects of many genes were dependent upon the environment, and that genes expressed in the parents affect offspring variation through transgenerational parental effects. While their ideas were incredibly useful to gain an understanding into how populations evolve, their rudimentary concepts of gene topology and function were a limitation to understanding how natural selection shapes genotypes into phenotypes. At the time of the formulation of the Modern Synthesis, evolutionary biologists could show that forms do change, and were shaped by natural selection, but almost no evidence was available for *how* forms evolved. The Modern Synthesis treated development as a "black box", through which genetic information somehow transformed embryos into complex organisms. It took many decades for advances in molecular biology and embryology to finally start untangling the processes of development, and how these can evolve, thus offering a view into how ontogeny held the key to evolutionary processes.

### **I.1.b. Deep homology, pleiotropy and *cis*-regulatory evolution**

With the advent of discoveries in modern Molecular and Developmental Biology, mechanistic biological processes could now begin to be integrated into evolutionary theory. Homology based approaches to identify genes across disparate taxa led to the finding that developmental genes, such as those encoding transcription factors (TFs) (including the *Hox* genes) and signalling molecules (such as Wnt ligands) were shared by many animals, and directed similar processes during development. Large genetic screens initially identified developmental genes in model organisms which were based on large effect mutations causing visible phenotypes that could be easily screened (e.g. Nüsslein-Volhard and Wieschaus, 1980). Amongst these were the *Hox* genes, as well as other homeodomain containing genes such as *Pax6*, the mutation of which resulted in eyeless *Drosophila* (Quiring *et al.*, 1994; Halder *et al.*, 1995), and whose mouse equivalent could rescue the eyeless phenotype (Halder *et al.*, 1995; Patrick Callaerts *et al.*, 1997). Likewise, researchers showed that the cnidarian *Achaete-Schute* homolog was able to induce the formation of sensory organs in *Drosophila* and form heterodimers with the endogenous *Drosophila* Daughterless protein, despite there being over 1 billion years of evolution separating these species (Grens *et al.*, 1995), outlining the incredible conservation of these genes across large evolutionary distances.

The discovery of the extent of this sharing of genetic information prompted Shubin, Taffin and Carroll to coin the term “deep homology” to describe this shared regulatory circuitry inherited over large evolutionary time involved in the development of often non-homologous structures (Shubin *et al.*, 2009), and suggested that evolution largely occurred through the tinkering of a conserved “genetic toolkit” (Jacob, 1977; Jacob, 1993)<sup>1</sup>. This deep conservation of genes led to an apparent paradox: How do species, which have such diversity in morphology and behaviour arise from a set of functionally similar genes? The answer to this question is still a much-debated issue in the field, but, as evidenced by embryological and developmental

---

<sup>1</sup> This idea in fact goes back to the discovery of the Lac Operon by Jacob and Monod, for which they eventually shared the Nobel prize in 1965 and led to Monod’s famous statement: “What is true for *E. coli* is also true for the elephant” (Monod, 1988).

studies, some explanation is likely due to changes in the precise spatio-temporal regulation of developmental genes.

By the end of the 20<sup>th</sup> century, there were few examples linking the evolution of a single locus to morphological or behavioural differences between species. Studies at that time began to show that the expression patterns of key developmental genes were often correlated with morphological differences between disparate taxa ( Warren *et al.*, 1994; Averof and Akam, 1995; Cohn and Tickle, 1999). Such studies were usually performed by comparing distantly related species, and suggested that changes in the expression patterns of developmental genes were a key contributor to macroevolutionary differences in morphology. However, there was little direct evidence showing how morphological evolution was shaped by differences between individual genes (i.e. what are the mutations necessary to cause evolutionary change).

A key aspect in formulating a theory of how conserved genes evolve came from the idea of pleiotropy (Carroll, 2008). Most genes involved in regulating developmental processes have multiple functions throughout development, and changes in the regions encoding these proteins are likely to have widespread deleterious effects on an animal's fitness, suggesting their coding sequences are largely under strong evolutionary constraints. For example, signalling molecules of the conserved Wnt signalling pathway are involved in several processes of segmentation and wing development in butterflies, and are the same genes that regulate specific aspects of wing patterning and pigmentation (Martin and Reed, 2014). Likewise, specific *Hox* genes - such as Ultrabithorax (*Ubx*) - have become integrated (co-opted) into new pathways involved in the specification of butterfly eyespots (Weatherbee *et al.*, 1999), without significant changes to the coding regions of these genes.

How then does evolution “teach old genes new tricks?” A large body of evidence has accumulated over the past decade and points to changes in the regulatory logic involved in directing the precise spatio-temporal expression of genes (e.g. Gompel *et al.*, 2005; Chan *et al.*, 2010; Stern and Frankel, 2013). Specific elements - often located a few hundred base pairs up or downstream of protein coding genes - called *cis*-regulatory elements (CREs), are able to bind



TFs that direct the expression or repression of downstream targets. Importantly, work dissecting the regulatory logic of CREs has shown that these elements are highly modular: arrays of CREs independently direct the expression of specific genes in different tissues and/or developmental contexts, and the altering of specific binding sites in CREs does not affect the pleiotropy of their downstream targets (Carroll, 2008). Moreover, under this *cis*-regulatory logic, a single TF can direct the expression of a myriad of downstream genes, the effect of which is often dependent on the *trans* environment (*i.e.* developmental context and developmental history) in which a particular TF is expressed (Stern, 2000). This modular action combined with the accumulation or loss of binding sites for specific TFs in CREs, allows gene regulatory networks (GRNs) to rapidly evolve without deleterious pleiotropic effects.

Numerous examples of *cis*-regulatory evolution both between and within species have been described. For example, in several *Drosophila* species, changes to the CREs controlling the expression of the pigmentation gene *yellow*, have been linked to variation in abdominal and wing pigmentation (Wittkopp, True, *et al.*, 2002; Wittkopp, Vaccaro, *et al.*, 2002; Werner *et al.*, 2010). Specifically, the wing spot of *D. biarmipes* evolved by the recruitment of a CRE which ancestrally functioned to give a low level of expression across the whole wing (Arnoult *et al.*, 2013). Following the co-option of transcription binding sites for Engrailed and Distalless, *yellow* became associated with expression in the epidermis prefiguring the wing spot. In another classical example of gene regulatory network co-option, the gene *wingless* (*wg*) has been linked to the evolution of melanic spots observed on the wings of *D. guttifera*, by the acquisition of binding sites responding to the *wg* morphogen at the *yellow* locus (Werner *et al.*, 2010). Mutations at specific binding sites between orthologous CREs have also been shown to underlie the recurrent evolution of divergent larval trichome patterns across several *Drosophila* species. The loss of trichome patterns in *D. sechellia* and the more distantly related *D. ezoana*, have been mapped to the same orthologous regions in CREs controlling the expression of *shavenbaby*, a “master regulator” involved in trichome formation (Stern and Frankel, 2013).

Deletions of CREs have also been shown to be important for the evolution of morphology. In populations of threespine stickleback fish (*Gasterosteus aculeatus*), recurrent deletions of the *Pituitary homeobox transcription factor 1 (Pitx1)* CRE, a gene involved in the development of pelvic spines, were shown to be responsible for the loss of these structures in several fresh-water populations (Chan *et al.*, 2010).

These examples, amongst many more, have highlighted the crucial aspect of mutations affecting the *expression* of downstream genes, rather than changes in the coding regions or changes to the number of genes. Over recent years there has been much debate with regards to the primary contributors to evolutionary change. Sean Carroll, a main proponent of the *cis*-regulatory hypothesis, argues that evolution must predominately occur through CRE evolution:

*“regulatory DNA is the predominant source of the genetic diversity that underlies morphological variation and evolution...”*

*“form evolves largely by altering the expression of functionally conserved proteins, and ... such changes largely occur through mutations in the cis-regulatory sequences of pleiotropic developmental regulatory loci and of the target genes within the vast networks they control”*

–Carroll, 2005, 2008.

While these findings describe some of the most remarkable examples of morphological evolution, they are by no means the only way morphology evolves. The logic behind *cis*-regulatory modulation is indeed intuitive and appealing, but perhaps the claim that evolution predominantly works through these principles is somewhat premature (Hoekstra and Coyne, 2007; Meiklejohn *et al.*, 2014). Several examples of changes to protein coding genes that affect animal form have indeed been described, including several instances of gene duplication and

divergence. The importance of gene duplications for morphological evolution has long been recognised, and is discussed in more detail below.

### **I.1.c. Gene duplication and divergence**

The fate of duplicated genes has been extensively studied since Sasumu Ohno first popularised their significance in the 1970s (Ohno, 1970). Gene duplications and their subsequent divergence play a key role in the evolution of novel gene functions. Evidence for whole genome duplications (WGDs) has been described in several species (Holland *et al.*, 1994; Flot *et al.*, 2013; Jaillon *et al.*, 2004; Dehal and Boore, 2005; Meyer and Van de Peer, 2005; Brunet *et al.*, 2006; Kenny *et al.*, 2016) as well as many instances of lineage specific gene expansions through tandem gene duplications (TGDs). Several other mechanisms such as retrotransposition, segmental duplication, *de novo* incorporation and exon copying also greatly influence the number of gene copies in animals' genomes. In fact, the number of protein coding genes often varies by many thousands between animal species (Holland *et al.*, 2017).

The mechanisms by which genes duplicate tend to be classified on the basis of their size, or whether they recruit an RNA intermediate. Reverse transcribed RNAs can be copied and integrated into random sites within a genome. Although these rarely give rise to expressed full-length coding sequences, several instances of functional and adaptive retrogenes have been described across disparate taxa (Long and Langley, 1993; Betrán *et al.*, 2002; Burki and Kaessmann, 2004). Segmental duplications (which are essentially large TGDs), can occur as a result of unequal crossing over, as a result of homologous recombination between regions of high similarity in the genome. Chromosomal positioning and sequence repeats can promote these cross-over events (Levinson and Gutman, 1987), and may result in genomic hotspots for gene duplication. Finally, the occurrence of WGDs through polyploidisation is arguably the most abrupt change a genome may experience during a single evolutionary occurrence.

WGD events have been described across a wide range of eukaryotic groups, including ciliates (Aury *et al.*, 2006), plants (Fawcett *et al.*, 2009; Li *et al.*, 2015), fungi (Wolfe and Shields, 1997; Ma *et al.*, 2009), oomycetes (Martens and Van de Peer, 2010) and animals

(Holland *et al.*, 1994; Jaillon *et al.*, 2004; Brunet *et al.*, 2006). Within the animal kingdom, WGDs have occurred at the base of the vertebrate lineage (Holland *et al.*, 1994; Dehal and Boore, 2005), in teleost fish (Amores *et al.*, 1998; Meyer and Van de Peer, 2005; Brunet *et al.*, 2006) as well as in salmonoid (Macqueen and Johnston, 2014) and cyprinid fish (Xu *et al.*, 2014) and likely contributed to organismal complexity associated with many vertebrates. Recently, studies have shown that WGDs may also have occurred up to twice in chelicerates (Kenny *et al.*, 2016; Schwager *et al.*, 2017). Studies of WGD reveal that, while gene number does not massively increase, developmentally important genes are preferentially retained, suggesting that these genes can be recruited to perform new and adaptive functions during subsequent radiations (Brunet *et al.*, 2006).

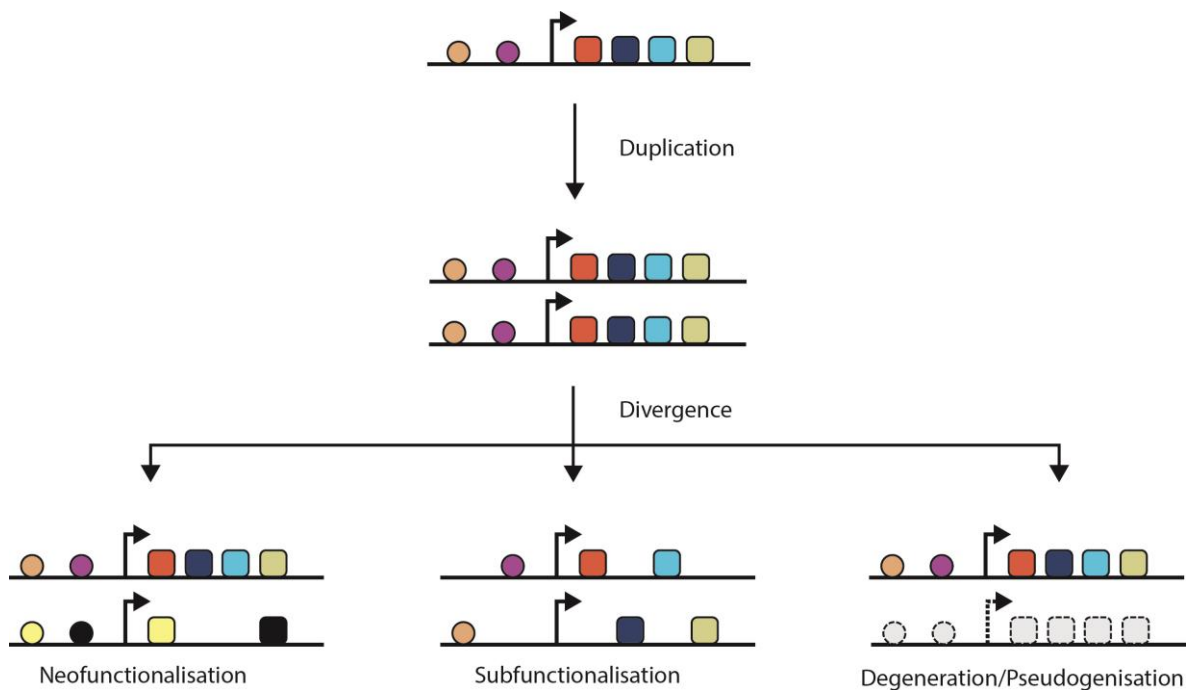
While WGDs are undoubtedly an important source of evolutionary innovation, they are relatively rare in comparison to TGD events. Many well described cases of TGD exist, and have been linked to lineage specific adaptive functions. For example, Lepidoptera display several lineage specific instances of gustatory receptor gene duplications, and have been argued to be a key innovation in the selection of hostplants, egg-laying sites and potential mates (Wanner and Robertson, 2008; Briscoe *et al.*, 2013). TGD events have also been described within lepidopteran and dragonfly opsin genes associated with novel, and often adaptive, ways of vision (Briscoe, 2001; Futahashi *et al.*, 2015) as well several instances of immune related genes associated with lineage specific pathogenic exposure (Tanaka *et al.*, 2008). In several of these cases, duplicate genes have substantially diverged from each other, suggesting individual cases of both neo- and subfunctionalisation following the duplication events.

Paralogs which arise through gene duplications are thought to be able to evolve rapidly due to a relaxed constraint, as the ancestral copy can maintain its original function, thus allowing new copies to diversify (Ohno, 1970; Innan and Kondrashov, 2010; Kondrashov, 2012). Following a gene duplication event, population genetic models show that for these new alleles to become fixed they must overcome substantial hurdles. Even for mutations offering considerable selective advantage, fixation is a relatively rare event, and estimates show that one

in a hundred genes is duplicated and fixed every million years (Lynch and Conery, 2000).

However, once duplications become fixed in a genome, three main fates for the duplicates have been described (Figure I – 1).

Duplicates are most likely to accumulate deleterious substitutions, either through the introduction of internal stop codons, or a mutation likely to affect the structure of a proteins' functional domain (Innan and Kondrashov, 2010), resulting in pseudogenisation or gene loss over time. However, it is possible that during the fate-determining phase of a duplicate, advantageous mutations become fixed in either coding or *cis*-regulatory regions leading to either neo- or subfunctionalisation. In the case of neofunctionalisation, accumulated mutations can confer a new expression domain and/or protein function, leading to one of the copies gaining a new role (Innan and Kondrashov, 2010). A further possibility for the gene duplicates is to subdivide the ancestral function (subfunctionalisation). Numerous genes work by performing many subtly distinct functions, and selective processes can result in partitioning of gene function after gene duplication. Partitioning these functions between the duplicates may in some cases increase the fitness of the organism by removing the conflict between two or more roles, effectively reducing pleiotropy (Makova and Li, 2003).



**Figure I - 1. Possible fates of duplicate genes.**

Consider a gene with four (pleiotropic) functions and two CREs (represented as coloured circles). Following a duplication event, a sister copy can acquire new functions through substitutions either within the coding region (squares) or CREs (circles), while the original maintains its ancestral function (neofunctionalisation). Sister copies may also subdivide (and potentially improve) the ancestral function (subfunctionalisation). Finally, duplicates may also degrade and/or stop being transcribed resulting in gene loss (degeneration/pseudogenisation).

In both cases of neo- and subfunctionalisation, patterns of accelerated rates of amino acid substitutions are often observed, leading to an increase in nonsynonymous over synonymous substitutions ( $d_N / d_S$ , or  $\omega$ ) over the duplicates' life (Yang and Bielawski, 2000; Nielsen, 2005). The idea is that by measuring the rate of substitutions between paralogs and orthologous sequences between species, positive selection can be measured as evolution favouring particular amino acid changes over time. On the other hand, sequences under purifying or negative selection will show an excess of synonymous substitution rates, as constraints on function penalise nonsynonymous substitutions that alter protein structure (Vitti *et al.*, 2013). The logic behind  $d_N / d_S$  rates is often used to infer episodes of negative or positive selection operating on a set of protein coding sequences, and has led to many discoveries of important positively selected genes and residues within proteins (e.g. Sawyer *et al.*, 2005; Brault

*et al.*, 2007; Soares *et al.*, 2008). Furthermore, sequence evolution following gene duplication is not limited to differences in paralogous sequences. Paralogs are also able to diversify following a speciation event, leading to asymmetric levels of divergence in *orthologous* sequences between species (reviewed in Holland *et al.*, 2017). It is possible that during speciation events, orthologs become adapted to species specific roles, and so divergence between orthologs can occur<sup>2</sup>.

Instances of increased  $\omega$  have indeed been detected in many cases following gene duplication events. For example, Lepidopteran opsin genes were shown to be under positive selection following a gene duplication event in the genus *Papilio* (Briscoe, 2001; Spaethe and Briscoe, 2004). Elevated rates of nonsynonymous substitutions were detected in a branch leading to a specific red-shifted photoreceptor post-duplication, in a transmembrane domain located in the chromophore-binding pocket of the visual pigment. Likewise, duplications of ionotropic receptors in *Heliconius* butterflies, which are involved in chemo-sensory perception, have experienced several rounds of positive episodic selection during various branches in their phylogeny (van Schooten *et al.*, 2016), and are argued to be key innovations involved in a wide range of behaviours. Similarly, the *Drosophila* CAF1-55 protein (a protein involved in the regulation of *Hox* genes through its interaction with the Polycomb repressive complex) was recently shown to have duplicated in the *obscura* subgroup, and subsequently experienced periods of asymmetric divergence from its sister copy (Calvo-Martín *et al.*, 2017).

Several instances of asymmetric gene divergence following gene duplication of homeobox genes, arguably some of the most conserved genes ever discovered, have also been described. For example, the hybrid-male sterility gene *Odysseus* (*OdsH*) shows a high degree of diversification both from its parental copy *unc-4* and between orthologous sequences in different *Drosophila* species (Ting *et al.*, 1998; Ting *et al.*, 2004). Loss of testis-specific

---

<sup>2</sup> It is important to note that sequence divergence between orthologous sequences does not - of course - necessarily result from a gene duplication event. Specific genes and individual residues which they encode can and do become targets of natural selection, but accelerated divergence is expected from duplicates as a result of constraint relaxation.

expression of *OdsH* causes male sterility in *D. melanogaster*. When the *D. mauritiana OdsH* allele is introgressed into *D. simulans*, misexpression of *OdsH<sup>mau</sup>* in testis causes hybrid male sterility (Ting *et al.*, 2004). The rapid sequence evolution is also in accordance with their divergent expression patterns, and is argued to have contributed to the speciation event between *D. mauritiana* and *D. simulans*.

Even within the *Hox* cluster, changes in structure and function of individual *Hox* genes have been linked to evolution of body plans. For example, loss of abdominal appendages in insects is explained by coding changes in both Abdominal-A and the Ubx protein, whereby the acquisition of a novel domain in Ubx allowed it to repress *distal-less (Dll)* (Averof and Akam, 1995; Averof and Patel, 1997; Ronshaugen *et al.*, 2002). *Dll* is a key gene involved in the development of appendages (Cohen *et al.*, 1989; Galant and Carroll, 2002; Grenier and Carroll, 2000). Likewise, in *Drosophila*, the *Hox* complex protein Fushi Tarazu (Ftz), lost its role as a homeotic gene, and acquired a new function as a pair-rule gene. Although this shift in function is partly explained due to changes in expression patterns through CRE modification, differences in protein function are also attributed to changes in coding regions, such as the acquisition of an LXXLL motif that is necessary for its functional interaction with a novel cofactor (reviewed in Pick, 2016). Furthermore, in squamates (lizards and snakes), *Hox* genes have been extensively modified in both their coding and *cis*-regulatory regions, which in part underpins their highly variable number of vertebrae (Di-Poï *et al.*, 2010). Marked increase in  $\omega$  ratios were reported for several posterior *Hox* genes, indicating that the coding regions were under significant relaxed selection during the stem-lineage of the squamates. Such relaxation is argued to have been a pre-requisite for the modifications that occurred to the snake axial skeleton, which resulted in a more anterior expression of *Hoxc6* in snakes, with the result of a homeotic shift from neck vertebrae to rib-bearing vertebrae and the loss of forelimbs (Woltering *et al.*, 2009).

Examples of diversification following gene duplication within the *Hox* cluster also exist, as exemplified by the *Drosophila bicoid* gene (*bcd*). During the radiation of the



Cyclorrhapha, a derived clade of Diptera, the homeobox gene *zerknüllt* (*zen*) duplicated through TGD to give rise to another copy of *zen* (*zen2*) and the highly divergent *bcd* (Stauber *et al.*, 1999; Stauber *et al.*, 2002). Ancestrally, *zen* is a gene involved in the specification of extraembryonic tissue during insect embryonic development (Schmidt-Ott *et al.*, 2010). Following mutations affecting key residues of the homeodomain, *bcd* became recruited to early axis specification, and now is the key gene sitting on top of the segmentation cascade in higher flies (Driever and Nüsslein-Volhard, 1988).

While TGD events in the *Hox* cluster are rare events in animal evolution, recurrent duplications of *zen* (known ancestrally as *Hox3*), have occurred several times independently during the radiation of the insects (Stauber *et al.*, 1999; van der Zee *et al.*, 2005; Schmidt-Ott *et al.*, 2010). In many of these reported duplications, the resulting paralogs have undergone high levels of divergence, and have acquired new and specific roles in the early development of individual species. Furthermore, the evolution of extraembryonic tissue is tightly linked to the evolution of the *Hox3/zen* gene as its transition from a canonical *Hox* gene to extraembryonic membranes coincides with the origin of these structures in insects (Schmidt-Ott *et al.*, 2010). It is argued that both this loss of canonical *Hox* function, and subsequent duplications, have allowed the development of these novel structures, and refinement thereof (Hughes and Kaufman, 2002; Schmidt-Ott *et al.*, 2010). Therefore, the case of *Hox3* presents an ideal opportunity to study TGDs and associated gene diversification patterns, allowing us to probe deeper into the contribution of recurrent duplications to the evolution of morphology.

#### **I.1.d. *Hox3* duplications and the Evolution of Extraembryonic tissue in Insects**

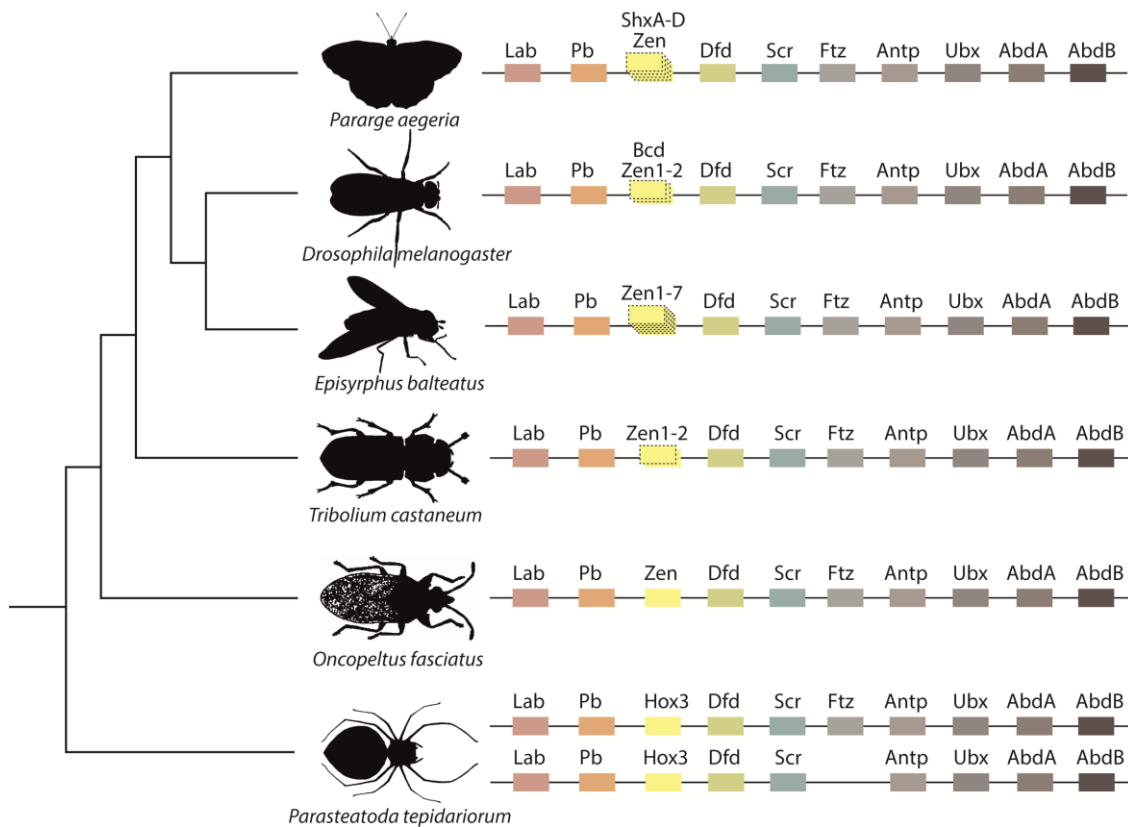
In insects, extraembryonic development starts at the blastoderm stage, during which a portion of cells is fate mapped to become the serosa, an epithelium located underneath the eggshell that covers the developing germ band or embryo proper (Schmidt-Ott *et al.*, 2010). Serosa formation usually occurs through the folding of dorsal tissue over the ventral blastoderm, or by invagination of posterior blastoderm (reviewed in Panfilio, 2008). Following these cell

rearrangements, the germ band becomes uncoupled from the serosal epithelium, as serosal cells completely envelop the embryo. In most pterygotes (winged insects), the serosa provides the outermost cellular epithelium, enclosing all other contents, including the amnion, embryo and yolk. The serosa is therefore the first layer of cellular membrane to be in contact with the environment, and has been argued to be a key adaptation that allowed insects to exploit a plethora of ecological niches, in part due to its protective function that buffers the embryo against external environmental perturbations and invasive assaults (Horn *et al.*, 2015). It is interesting to note that the evolution of extraembryonic epithelium accompanies, in large part, the terrestrialisation of arthropods (Jacobs *et al.*, 2013). While terrestrial arthropods such as chelicerates (which do not have a true serosa) are very well adapted to dry environments, their eggs are usually protected through extensive parental care and the covering of embryos through cocoons (Mittmann and Wolff, 2012). Likewise, Myriapods, crustaceans and Entognatha are generally found in humid environments, and their extraembryonic membranes do not entirely cover the embryos (Jacobs *et al.*, 2013). In contrast, the vast majority of insects are completely enveloped with an active serosal layer, and recent evidence suggests it provides a protective function against desiccation (Jacobs *et al.*, 2013; Horn *et al.*, 2015). For example, recent studies have shown that the serosa is implicated in desiccation/drought resistance, through the secretion of a cuticle which also provides mechanical support to the egg itself (Jacobs *et al.*, 2013; Jacobs *et al.*, 2014). Furthermore, the serosa is able to provide a direct defence from external exposures through the processing of environmental toxins and by mounting immune responses in a range of insects, including beetles and the ditrysian moth *Manduca sexta* (Jacobs *et al.*, 2013; Jacobs *et al.*, 2014; Lamer and Dorn, 2001; Berger-Twelbeck *et al.*, 2003; Orth *et al.*, 2003; Rezende *et al.*, 2008). For example, experiments performed in the beetle *T. castaneum* where serosal-less embryos were exposed to varying levels of humidity (Jacobs *et al.*, 2013), and pathogens (Jacobs *et al.*, 2014), have shown that the serosa is not only required for embryos to survive under low humidity environments, but also to mount an adequate immune response following (invasive) pathogenic exposure.

Extraembryonic tissue evolution is tightly linked to evolution of the *Hox3/zen* transcription factor. During the early radiation of the insects, the ancestral *Hox3* lost its role in specifying segments along an anterior-posterior axis and attained a role in the specification of extraembryonic tissue (Schmidt-Ott, 2010). Comparative expression data across different arthropod lineages suggests that this evolutionary transition may have arisen during the radiation of the pterygota (Hughes *et al.*, 2004). Subsequently, the *Hox3* locus has been shown to have duplicated independently in a variety of insects including the beetle *T. castaneum* to yield two copies named *zen1* and *zen2* (after the *Drosophila* phenotype), in the hoverfly *Episyrphus balteatus* to give seven copies of *zen* (Rafiqi *et al.*, 2008) and in the fruit fly *D. melanogaster* to give rise to two copies of *zen* and the highly divergent *bcd* (Stauber *et al.*, 1999; Stauber *et al.*, 2002). More recently, it was shown that a further round of duplications occurred during the radiation of the Ditrysia, a derived clade of the Lepidoptera, to give rise to 5 highly divergent *Hox3* genes named the *Special Homeobox (Shx)* genes (Figure I – 2) (Ferguson *et al.*, 2014)<sup>3</sup>.

---

<sup>3</sup> Evidence of *Hox3* duplication also exists outside of the insects, in the centipede *Strigamia maritima* and both appear *Hox*-like in sequence (Chipman *et al.*, 2014). Interestingly, both copies seem to be found outside of the *Hox* cluster.



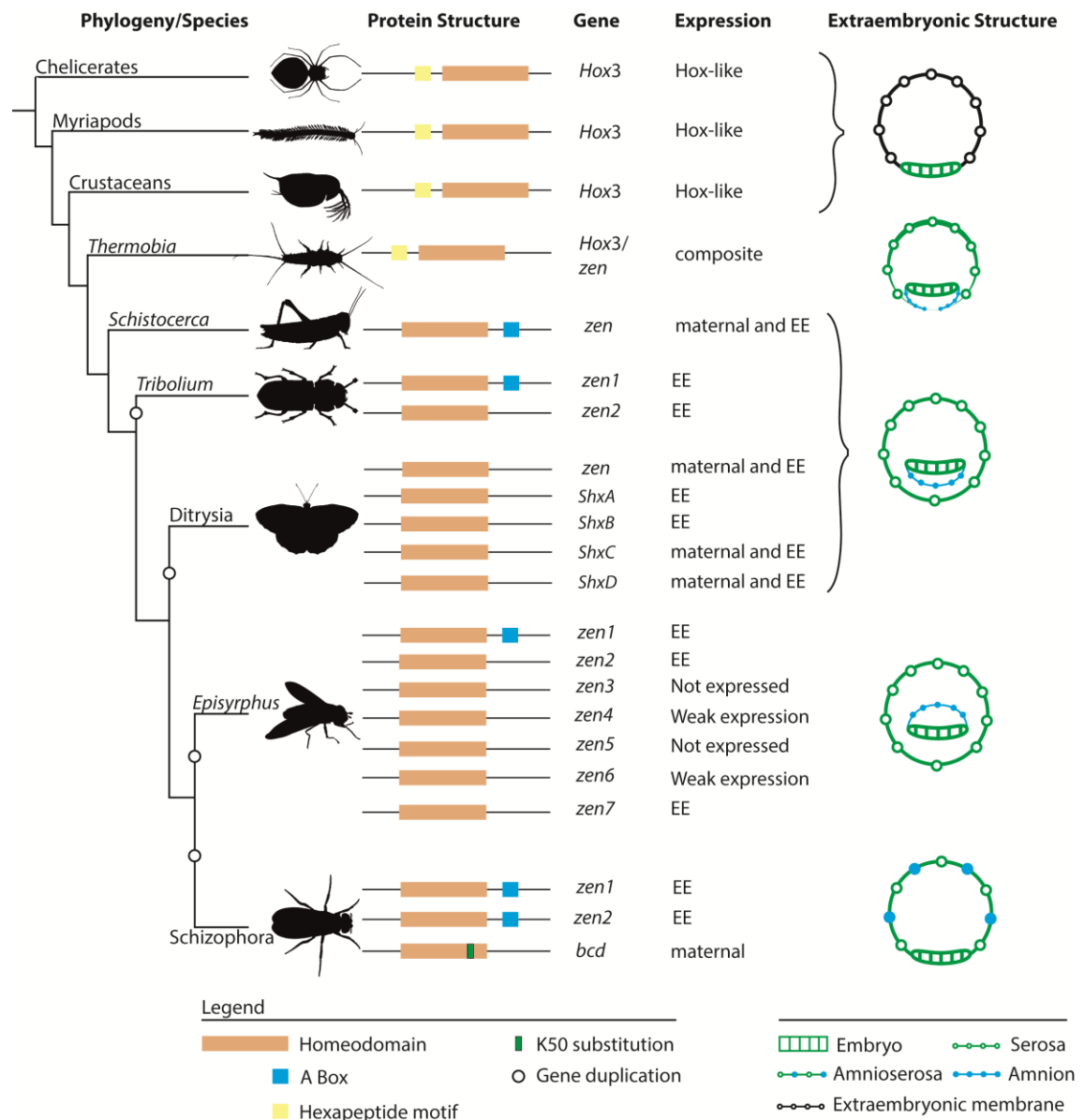
**Figure I - 2. Hox cluster expansion in arthropods.**

Ancestrally, *Hox3* functioned as a canonical *Hox* gene in arthropods such as chelicerates. During the radiation of the pterygota, *Hox3* then became recruited to the development of extramembryonic tissue, and is known as *zen*. *Zen* has duplicated various times in independent insect lineages, giving rise to two copies in the beetle *T. castaneum*, to two copies of *zen* and the highly divergent *bcd* in *D. melanogaster* and seven copies in *E. balteatus*. *Zen* duplicated to give rise to four extra copies of *Hox3* paralogs in Ditrysia, named *ShxA* to *ShxD*. Note that chelicerate duplication of *Hox3* is a result of WGD not TGD.

While there is evidence that *cis*-regulatory changes are also associated with the divergence in *Hox3* function during insect evolution, there is also substantial evidence that coding changes played a major role in both the transition of *Hox3* to *zen*, and in the subsequent diversification of duplicates (Figure I- 3). In ancestral arthropods, *Hox3* is expressed as a canonical *Hox* gene, and is involved in morphological differentiation of segment identity (Hughes *et al.*, 2004; Panfilio and Akam, 2007; Papillon and Telford, 2007). *In situ* hybridisations in the spider *Cupiennius salei* show that it's *Hox3* ortholog is expressed in regions corresponding to the pedipalp segment, and in the four leg bearing segments (Damen and Tautz, 1998). Likewise, in the spider *Parasteatoda tepidariorum*, both *Hox3* copies (which arose from WGD), retained a *Hox*-like expression pattern (Schwager *et al.*, 2017). In the

centipede *Lithobius atkinsoni*, *Hox3* expression is limited to the intercalary and mandibular segments (Hughes and Kaufman, 2002), and *Hox*-like expression is also retained in the crustacean *Daphnia pulex* (Papillon and Telford, 2007).

Remarkably, the *Hox3* ortholog in the basal apterygote insect *Thermobia domestica*, has an intermediate expression pattern, where it is localised to both extraembryonic tissue and in the mouth bearing segments (Hughes *et al.*, 2004). This transition from *Hox*-like to extraembryonic is also accompanied by several protein structure rearrangements. In *T. domestica*, the transition reflects a shift of the homeodomain towards the amino terminal, although its sequence clusters more readily with *Hox3* in basal arthropods (Hughes *et al.*, 2004; Panfilio and Akam, 2007). Following the radiation in the pterygotes, sequence analysis shows further migration of the homeodomain towards the amino terminal, a gain of a conserved “A-box motif”, and a loss of the hexapeptide (YPWM) motif, which enables interaction of *Hox* genes with the *Hox*-cofactor Exd/Pbx (Mann and Chan, 1996; Panfilio and Akam, 2007). In the context of *Hox* gene expression, this co-factor is crucial to confer DNA binding specificity to the individual *Hox* genes (Mann and Chan, 1996). At the base of the pterygotes, this transition is further pronounced. In the grasshopper *Schistocerca gregaria*, the *Hox3* ortholog (which after the radiation of the pterygotes is now referred to as *zen*), has a further shift of the homeodomain towards the amino terminal, and a complete loss of the hexapeptide motif (Dearden *et al.*, 2000). Expression now is also exclusive to extraembryonic tissue, and in *S. gregaria*, attained a novel maternal expression, where it is also expressed in the oocyte cytoplasm, and delineates the boundary between amnion and serosa (Dearden *et al.*, 2000).



**Figure I - 3. Duplications and functional divergence in *Hox3* genes across the arthropod phylogeny.**

In ancestral arthropods such as Chelicerates, Myriapods and Crustaceans, *Hox3* is expressed as a canonical *Hox* gene, establishing the identity of segments along the AP axis. These arthropods are also characterised by the absence of serosal and amniotic tissues. In apterygotes, *Hox3* expression became associated with extraembryonic (EE) membranes, but also retained a *Hox*-like expression pattern. Following the radiation of the pterygotes, *zen* expression became associated exclusively with extraembryonic membranes, and lost a hexapeptide motif required for *Hox* gene co-factor interaction. The loss of this motif is accompanied by the gain of a conserved motif outside of the homeodomain called the A-box, as well as a shift of the homeodomain towards the amino terminal. Subsequently, *zen* became duplicated in Coleoptera, Ditrysia and higher flies independently, leading to cases of both neo- and subfunctionalisation. EE; Extraembryonic. See text for further details and relevant references. Style of depiction of embryonic and extraembryonic tissue is after Jacobs *et al.*, (2013).

In the beetle *T. castaneum*, *zen* has become duplicated and acquired new functions in dorsal closure (the process by which the embryonic midline becomes fused during embryogenesis), while the original copy maintained its role in serosal specification (van der Zee *et al.*, 2005). The distinction in these functions is argued to be a direct consequence as a result of subfunctionalisation of the *zen* paralogs following duplication. Both copies are expressed throughout the serosal tissue, but only *zen2* is expressed in the developing amnion (van der Zee *et al.*, 2005). Knockdown of *zen1* by RNAi, remarkably, produces viable embryos and abolishes serosal development that results in the development of a single epithelium, in which all cells appear amniotic (Jacobs *et al.*, 2013), whereas knockdown of *zen2* abolishes the fusion of the amnion and serosa and prevents amniotic rupture during katanepsis (the process by which the embryo inverts in the egg, re-fusing with extraembryonic membranes before absorbing them) (Panfilio, 2008; Horn *et al.*, 2015; Schmidt-Ott and Kwan, 2016).

At the base of the cyclorraphan Dipterans, *zen* further duplicated independently to give rise to an extra copy of *zen*, as well as the highly derived *bcd* (Stauber *et al.*, 1999; Stauber *et al.*, 2002). Following this duplication, *bcd* evolved extremely rapidly, and gained new maternal expression, as well as several key substitutions which allowed it to recognise different downstream targets (Stauber *et al.*, 1999). These include a key lysine substitution at position 50 (K50) conferring new target recognition, and an arginine at position 54 which conferred new RNA binding capabilities (Hanes and Brent, 1989; Treisman *et al.*, 1989). Maternal *bcd* mRNA is localised to the anterior pole of *Drosophila* embryos, where it forms a protein gradient with a maximum concentration at the anterior pole (Driever and Nüsslein-Volhard, 1988). The Bcd protein functions as a morphogen, where downstream targets respond in a concentration dependent manner to allow the formation of segments (Berleth *et al.*, 1988). Knockdown of *bcd* results in embryos lacking a head and thorax, but does not directly affect the development of extraembryonic tissue (Driever and Nüsslein-Volhard, 1988). Thus, during the diversification of *bcd*, it acquired a role as an anterior determinant, and lost any similarity in function with its progenitor gene *zen*.

In terms of extraembryonic specification, higher dipterans (specifically the Schizophora) are also very unusual, as they display a reduced extraembryonic tissue, where serosa and amnion have become fused (amnio-serosa) (Schmidt-Ott, 2000; Schmidt-Ott *et al.*, 2010). While the two *zen* copies appear to be dispensable for extraembryonic development in *Drosophila* (Rushlow and Levine, 1990), a loss of postgastrular expression of *zen1* is argued to be the cause for the formation of the single, fused amnio-serosa (Rafiqi, 2008). In the basal cyclorraphan fly *Megaselia abdita*, postgastrular suppression of *zen* results in the failure of serosal tissue to disjoin from amniotic cells, resulting in an amnioserosal-like layer (Rafiqi, 2008). Therefore, the loss of later *zen* expression in *Drosophila* may have resulted in the establishment of a single, fused amnio-serosal layer. The reduction in extraembryonic epithelia also accompanies a shift in general egg laying strategies in higher flies. Schizophoran flies generally lay their eggs in rotting vegetable matter or damp soil, or in the case of desert adapted flies, eggs are deposited in the necrotic tissue of cacti, where relative humidity can reach 90% at night (Gibbs *et al.*, 2003). It is therefore possible that Schizophoran flies may be able to dispense with the protective function of serosal membranes covering the entire embryo during embryogenesis. Interestingly, an expansion of the *Hox3* locus resulting in seven *zen*-like genes was also reported in a non-Schizophoran Dipteran, the hoverfly *E. balteatus* (Rafiqi, 2008). Five of these genes were found to be expressed during early development, three of which were shown to directly contribute to the specification of serosal tissue (Rafiqi, 2008). RNAi experiments revealed serosal defects only for *zen-2* and *zen-7* in addition to the original *zen* gene, suggesting that the other paralogs might be dispensable for serosal and embryonic specification or may have become pseudogenised. Analysis of the homeodomain of two of the paralogs indeed suggests that this is the case, as they contain large deletions. Moreover, the expression patterns and RNAi phenotypes of *zen-2* and *zen-7* in *E. balteatus* hint at differing functions, suggesting the paralogs might have also been neo- or subfunctionalised in this species<sup>4</sup>.

---

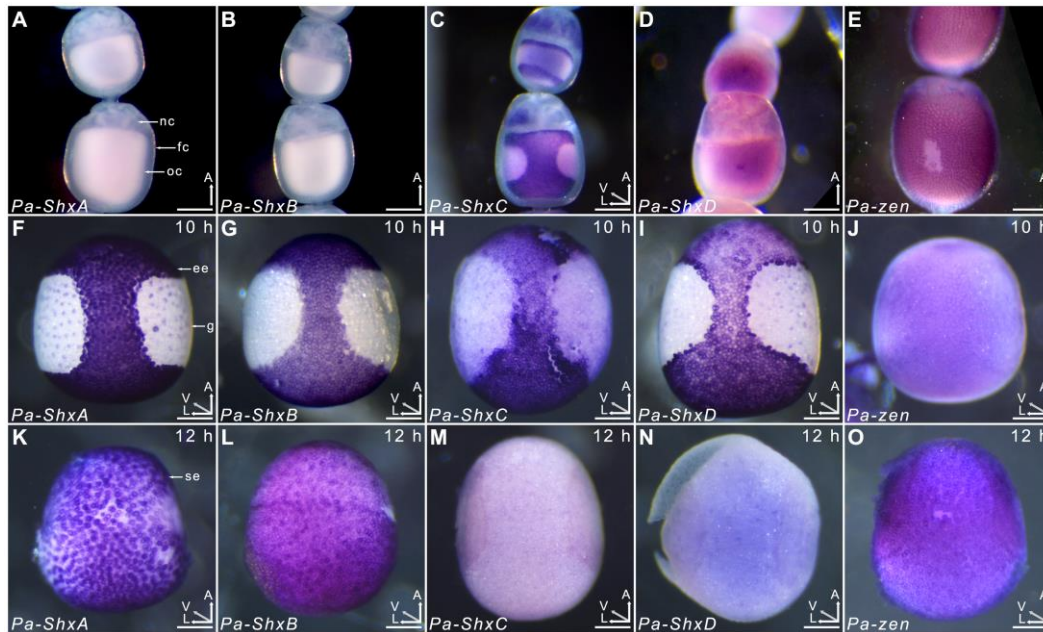
<sup>4</sup> Note that amino acid alignments of the *zen* genes in *E. balteatus* only recovers a putative A-box motif for *zen-1*, while it appears to be lost in subsequent duplicates.



More recently, a further expansion of the *Hox3* locus was identified in the Lepidoptera (Chai *et al.*, 2008; Ferguson *et al.*, 2014; Holland *et al.*, 2017). An initial screen of the *B. mori* genome identified a cluster of genes located in the *Hox* cluster between *proboscipedia* and *deformed*, which were named the *Shx* genes (Chai *et al.*, 2008). Analysis across a range of Lepidopteran species revealed that the *Shx* genes arose as a result of TGD of *zen* during the radiation of the Ditrysia and that most species retained a core set of four copies of *Shx* (*ShxA-D*) alongside *zen* (Ferguson *et al.*, 2014; Chapter II)<sup>5</sup>. Expression patterns recovered for the 5 *Hox3* genes in the speckled wood *Pararge aegeria*, revealed localised expression for the all four *Shx* genes to the serosa as well as differing expression patterns after serosal formation over the first 48 hours of embryogenic development (Ferguson *et al.*, 2014) (Figure I - 4). Three of the transcripts, *zen*, *ShxC* and *ShxD*, are maternally loaded into *P. aegeria* eggs. Transcripts of *ShxC* show localisation in the nurse cells and within the oocytes, where a striking “hour-glass” pattern is observed. This pattern pre-figures the region fated to become extraembryonic tissue, and corresponds to the presumptive serosa. *ShxD* transcripts are present throughout the developing oocyte, while *zen* transcripts are detected in the follicle cells surrounding the oocyte (Ferguson *et al.*, 2014; Carter, 2014). At the onset of zygotic transcription (~ 8h AEL), all *Shx* genes are detected in a clear hour-glass pattern in the cellularised blastoderm, resulting in a clear boundary between future extraembryonic cells and embryonic regions. In contrast, *zen* transcripts are only weakly localised throughout the blastoderm. At this stage extraembryonic cells have started differentiating, and appear larger and often polyploid. In *P. aegeria*, and most other butterflies (Braak *et al.*, in prep, Kobayashi *et al.*, 2003), serosal cells then migrate over the germ band, as the embryo and associated amniotic cells “sink” into the egg and become completely encapsulated by the serosa (Figure I – 5). Interestingly, and in contrast to most other insects, the serosa remains intact until hatching.

---

<sup>5</sup> While these genes are likely direct duplicates of *zen*, and could therefore be named *zen-2* to *zen-5*, I adopt the nomenclature of Chai *et al.* (2008) and Ferguson *et al.* (2014) to outline their divergent expression patterns and rapid sequence evolution.

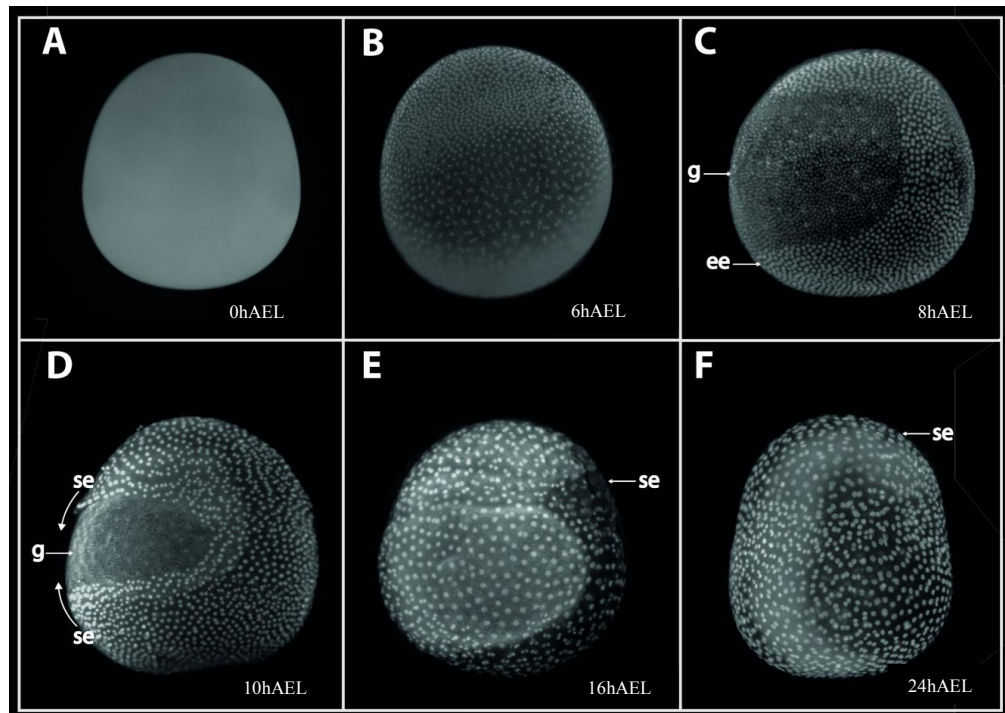


**Figure I - 4. Localisation of *Shx* transcripts in *Pararge aegeria* ovarioles (A-E), 10h embryos (F-J) and 12h embryos (K-O).**

Embryos and oocytes are orientated with the anterior to the top. Embryos dorsal side facing while lower and upper oocytes in C show dorsal and ventral faces respectively. Note that in 12 h embryos the serosal cells have migrated over the germ anlage forming an enveloping layer. Some follicle cells in E are removed to show absence of staining in the oocyte. Labels indicate nurse cells (nc), follicle cells (fc), oocyte (oc), germ anlage (g), and extra embryonic anlage (ee) which differentiates into the serosa (s). Orientation for each panel is indicated in bottom right 3D axis indicating anterior (A), left (L) and ventral (V) when known. AEL, after egg-laying (hours). Scale bars 200  $\mu$ m. Reproduced from Ferguson *et al.*, (2014).

The duplication of *Hox3* in Lepidoptera appears to be accompanied by a striking level of sequence divergence, even within the homeodomain (Ferguson *et al.*, 2014 and see Chapter II). Phylogenetic analyses argue for fast rates of evolution associated with the duplicates (long branch lengths), as well as extensive divergence between orthologous sequences in different species, something that has been studied extensively in Chapter II (with relevant references therein). There is evidence to suggest that even small changes at the level of the homeodomain can have large consequences on the ability of TFs to recognise different targets, as exemplified by the K50 substitution in the *bcd* homeodomain (Hanes and Brent, 1989). A key question that arises from the divergence between the *Shx* genes is which evolutionary forces (i.e. drift, positive and/or negatively selection) have shaped the variation observed. (See Chapter II and Chapter III). Specifically, is it possible to associate periods of positive selection following the

duplication of the paralogs, and between orthologous sequences in different species? And what roles might the *Shx* genes be playing that are specific to the Lepidoptera, and how significant are they in their massive radiation and diversification of life-styles?



**Figure I - 5. Early embryonic development in *Pararge aegeria*.**

A freshly laid egg is shown in A (i.e. 0hrs after egg-laying, AEL). A single nucleus is present at this stage (not visible here) which migrates to the egg surface. Cleavage proceeds in a syncytial blastoderm (B), from the anterior pole of the egg towards the posterior. Asynchronous rounds of cell division can be observed as cells migrate towards the posterior pole of the egg. Blastoderm cellularisation then occurs, following by early cell differentiation (C). Large polyploid extraembryonic cells become evident in a horseshoe pattern. Germ band cells appear smaller and more tightly spaced. As serosal cells differentiate further, they begin migrating over the germ band (D). Serosal cells then completely envelop the embryo (E), as the germ band continues differentiation below (F). Abbreviations: g – germ band; ee – extraembryonic; se - serosa. Anterior facing up, posterior facing down. Embryos have been dechorionated and the vitellin membrane has been removed for clarity. Braak et al, in prep.

Furthermore, levels of standing variation associated with TFs have the potential to be selected upon within populations, and are what natural selection shapes to eventually give rise to intraspecific differences (e.g. between populations experiencing significantly different environmental factors). Much of what is known about gene duplication and divergence is

inferred from studies on interspecific sequence variation in paralogs, often across a large phylogenetic scale, and thus not much is known about standing intraspecific paralog variation and associated phenotypic variability, nor how it may be of significance in local adaptation across a range of populations occupying widely divergent habitats. Quantifying the segregating variation within populations at TF loci is necessary to infer selective pressures and to ascertain the functional effects of naturally occurring allelic variation and sequence divergence among paralogs. This is one of the main aims of this thesis, and to quantify variation in the *Shx* genes across a large phylogenetic scale, within a genus (*Heliconius*), as well as within a single species (*Pararge aegeria*) (Chapter II; see also Section I.2). The few studies that have examined segregating variation among TFs have shown that polymorphisms are maintained by balancing and, in some cases, there is significant positive selection (Balakirev and Ayala, 2004; Jovelín *et al.*, 2009; Balakirev *et al.*, 2011).

Studying the demographic and population history of a species is a key aspect in understanding the factors influencing the ability of populations to adapt to local environments. Genetic variation at both the geographic and population level holds footprints of historical demographic events as well as existing processes such as the degree of gene flow and dispersal events between populations. It is therefore important to study the historical biogeographical events of a species in order to disentangle signatures of population structure from that of selection in genetic data. Chapter III will therefore detail the patterns of intraspecific *Shx* gene variation identified in *P. aegeria* in Chapter II within such a historical biogeographical context.

### **I.1.e The Speckled Wood Butterfly, *Pararge aegeria***

*Pararge aegeria*, the Speckled Wood butterfly, is a temperate butterfly species, and a popular model species for studies on plasticity in female reproduction (Gibbs and Dyck, 2009; Gibbs *et al.*, 2010a; Gibbs *et al.*, 2010b), life-history evolution (Nylin *et al.*, 1995; Gotthard and Berger, 2010), thermoregulatory behaviour (Van Dyck and Wiklund, 2002; Kemp *et al.*, 2006), development time (Sibly *et al.*, 1997), and morphology (Van Dyck and Wiklund, 2002; Merckx and Van Dyck, 2006; Breuker *et al.*, 2007, 2010). Apart from such evolutionary ecological

studies, the species has recently been successfully developed as a system to study maternal regulation of early embryogenesis and embryogenesis itself (Carter et al 2013, 2015, Ferguson et al 2014). *Pararge aegeria* has a widespread distribution, experiencing a large variety of environmental conditions, from cold and wet conditions in the North of Europe to hot and dry in the South of Europe and North Africa (Weingartner et al., 2006; Habel et al., 2013; Tison et al., 2014). Life-history traits vary widely between different populations, including direct development through larval and pupal stages into adults within the same season, winter diapause at the larval stage as well as diapause at the pupal stage (Aalberg Haugen and Gotthard, 2015). Such species have the potential to reveal broad biogeographical patterns associated with responses to both biotic and abiotic factors, thus providing valuable data on how species react to environmental pressures across large latitudinal ranges. In particular, understanding the way species respond to climatic conditions is of major interest as these can have large effects on a species ability to adapt to changing environments (Parmesan et al., 1999; Fisher et al., 2010). Furthermore, very little is known about how Lepidoptera (or most insects) are able to cope with this variety of environmental conditions at the embryonic stage. *Pararge aegeria* embryos are able to survive at a large range of temperatures and humidity, but some European populations do struggle at low humidity (egg hatching rate 50% at 20% RH, Braak et al, unpublished), particularly at high temperatures (e.g. a egg hatching rate of 20% at 27 °C and 23% RH has been observed for Madeiran *P. aegeria*, Gibbs et al., 2010). Given that the extraembryonic membranes appear to be involved in conferring a protective function to the embryo, a major question is how species that are adapted to such disparate environments manage to regulate early extraembryonic development, and whether different populations do so differently.

Several genomic resources have recently been established for *P. aegeria*, including an RNA and miRNA ovarian transcriptome (Carter *et al.*, 2013; Quah *et al.*, 2015), a pipeline to analyse high-throughput expression data (Carter *et al.*, 2016), as well as a draft assembly of a genome based on next-gen illumina sequencing (Ferguson *et al.*, 2014) and a complete mitochondrial genome (Teixeira da Costa, 2016). Moreover, it remains one of the few

Lepidopteran species for which early developmental tools are available, including established protocols for whole mount *in situ* hybridisation (Ferguson *et al.*, 2014; Carter *et al.*, 2015), and a detailed description of both oogenesis (Carter *et al.*, 2013) and embryogenesis (Braak *et al.*, in prep). Finally, it is in *P. aegeria* that it was shown, for first time in Lepidoptera, how extraembryonic tissue is specified by the mother, and that the expression patterns of all *Shx* genes is involved in patterning the serosa (Ferguson *et al.*, 2014). The Speckled Wood butterfly is therefore an ideal species with which questions relating to extraembryonic development, function and diversity can be addressed.

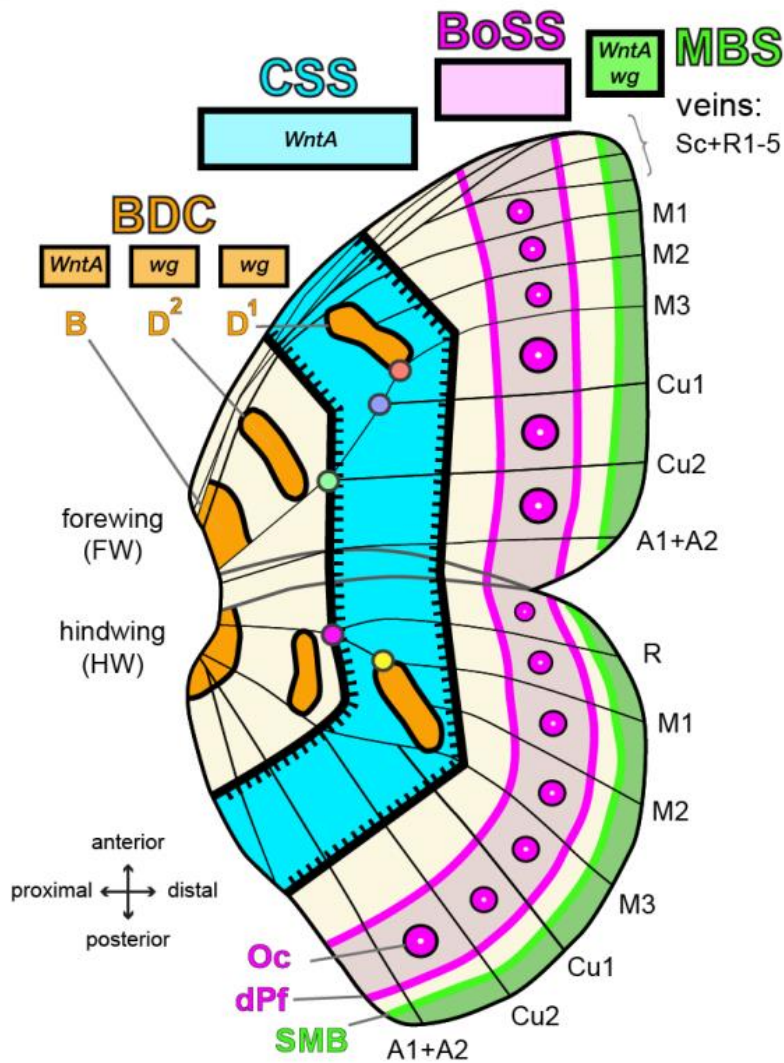
### **I.1.f Development of gene editing tools in *Pararge aegeria***

Despite Lepidoptera being the third largest insect order, far less is known about their early embryonic development than most other insect orders. Recent work on *P. aegeria* has shown this in all likelihood to be quite divergent with respect to other insect orders (Carter *et al.*, 2013, 2015). This disparity in data availability has in part been hindered by the relatively slow progress in developing functional tools for this order. While some progress has been made in species such as *Bombyx mori* (Takasu *et al.*, 2010; Takasu *et al.*, 2013; Takasu *et al.*, 2016) and *Bicyclus anynana* (Chen *et al.*, 2011; Marcus *et al.*, 2004; Monteiro *et al.*, 2013; Ramos and Monteiro, 2007; Ramos *et al.*, 2006; Tong *et al.*, 2014), there remains a need for the development of tools which can be applied to a variety of species, in order to start unravelling gene function across the Lepidoptera, particularly with a view to unravel how developmental evolution underpins the morphological diversity witnessed in this order.

Several genome editing techniques are available that enable the study of gene function, but perhaps the one showing the most promise is the recently developed clustered regularly interspaced short palindromic repeat/CRISPR-associated (CRISPR/Cas9) system. Application of the CRISPR/Cas9 technique has been successful in a wide range of species, including several Lepidoptera (Fujiwara and Nishikawa, 2016; Markert *et al.*, 2016; Perry *et al.*, 2016; Zhang and Reed, 2016; Zhang *et al.*, 2017; Mazo-vargas *et al.*, 2017). With this in mind, I developed the technique for my model species *P. aegeria*, which is reported in detail in Chapter IV. In the first

instance, key genes were targeted involved in wing patterning GRNs, as these provided mutant phenotypes that were relatively easy to ascertain, and thus to validate the technique in *P. aegeria*. The two key genes were the *yellow* gene, involved in the biosynthesis of wing pigments (Wittkopp and Beldade, 2009), and *WntA*; a gene which has been repeatedly mapped as the causative locus for wing patterning differences across a wide range of butterflies (Belleghem *et al.*, 2017; Gallant *et al.*, 2014; Jiggins *et al.*, 2017; Kronforst and Papa, 2015; Martin and Reed, 2014; Mazo-vargas *et al.*, 2017).

Across the nymphalid butterflies, the existence of a conserved “ground plan” that underlies the diversity of wing patterning has been proposed (Figure I – 6) (Schwanwitsch, 1924; Nijhout, 1978). Under this model, symmetric, homologous systems present across the wing surface are presumably altered during the course of evolution, to produce the variety of wing patterns observed in nature. Many of these patterns are proposed to arise from localised morphogenetic sources, which diffuse across the wing surface, producing symmetrical patterns surrounding that source (Nijhout *et al.*, 2003; Otaki, 2011). While wide ranging evidence has been shown for this hypothesis (Nijhout *et al.*, 2003), no study to date has directly tested the contribution of individual genes across different species through direct knockdown/knockout.



**Figure I - 6. The nymphalid groundplan.**

Consecutive symmetry systems organised along the antero-posterior axis are shown. These are generally categorised into the Basal system (BDC), containing the Basal (B), and discal first element (D1); Central symmetry system (CSS), containing the second discal element (D2); Border ocelli symmetry system (BoSS) containing the ocelli (Oc) and distal parafoveal element (dPf); and marginal band system (MBS). Expression of key morphogens hypothesised to be the source of patterns are shown (*WntA* and *wg*). Figure reproduced from Mazo-vargas *et al.*, 2017.

Further to the wing patterning genes, with their easily identifiable mutant phenotypes as a control, the main aim in Chapter IV was to target the *Hox3* paralogs in *P. aegeria*, in order to study the contribution of the *Shx* genes to serosal development, and early embryogenic development in general. While expression patterns recovered for the 5 *Hox3* genes strongly



argue for their role in the specification and maintenance of extraembryonic tissue (Ferguson *et al.*, 2014), no direct evidence is yet available for their function as serosal inducers, or for that matter any other developmental role they may play (i.e. pleiotropic effects). A key question with regards to the *Shx* genes is to what extent has each of the duplicates acquired new roles, or subfunctionalised ancestral functions? Disentangling the relative contributions of the paralogs through functional analysis is essential to gain an understanding to these questions.

## **I. 2. Aims**

Using *P. aegeria* as the model species, the primary aim of this thesis is to examine the significance of the recently duplicated *Hox3* genes for Lepidopteran development and in general their biology. A key aspect in addressing this question is to show that the variation associated with the *Shx* genes is likely functional and has been shaped by natural selection. Furthermore, given that the serosa is implicated in several functions involved in buffering the embryo to environmental perturbations, I was interested in examining polymorphisms at the population level that could be responding to environmental pressures. Finally, the direct contribution of the *Shx* genes to serosal development was investigated by means of knockout experiments.

### **I.2.a. Research objective 1 - Interspecific divergence and selection on *Hox3* locus across the Lepidoptera**

Given the large variation observed between paralogous and orthologous *Shx* sequences, the main aim in Chapter II was to measure selection operating at the *Hox3* locus at 3 different scales: across a large phylogenetic scale (and thus a wide range of Lepidoptera), within a genus, and within a species. For interspecific data, I measured levels of  $\omega$  across distantly related Lepidoptera, as well as within the more closely related *Heliconius* group. Population genetic statistics were applied to intraspecific data.

### **I.2.b. Research objective 2 – Geographical distribution of polymorphisms associated with the *Hox3* locus in *Pararge aegeria***

Using *P. aegeria* as a model system, in Chapter III, I examine the phylogeographical distribution of the speckled wood butterfly across Europe. This was done by sequencing the cytochrome oxidase subunit I (*COI*) and the developmental gene *wg* as a nuclear marker (cf. Weingartner *et al.*, 2006). The historical biogeography was used to compare against the spatial clustering of polymorphisms associated with the *Hox3* genes, which were sequenced for a subset of individuals across the same geographic range. Patterns of polymorphisms in the *Hox3* genes were used to infer potential episodes of selection operating at the locus.

### **I.2.c. Research objective 3 – Development of CRISPR/Cas9 technology in *Pararge aegeria*.**

Through injection of guide RNAs targeting the wing patterning genes *yellow* and *WntA*, as well as the *Hox3* genes *ShxA* and *ShxC*, I aimed to introduce targeted mutations at these loci. Mutations affecting the pigmentation genes are expected to result in patterning defects across the wings of *P. aegeria*, thus offering an amenable system to track the effect of the CRISPR/Cas9 injections. Following the establishment of the technique, I aimed to introduce mutations at the *Shx* genes, in order to study their contribution to serosal specification and maintenance.

## Chapter II

### *Hox3* polymorphism and signatures of selection in Lepidoptera

## Abstract

Paralogs arise through gene duplications and their subsequent divergence provides the raw material for functional innovation. Paralogs in the *Hox* cluster are rare, but the paralogy group 3 (PG3) has undergone independent tandem duplications in several insect clades. Within the Ditrysia, a derived lepidopteran clade, duplications in the ancestral PG3 gene *zerknüllt* (*zen*) resulted in 4 so-called special homeobox genes (*Shx*). It has recently been shown that these genes display great sequence divergence at the macro-evolutionary level, even within the highly conserved homeodomain, but at present we do not know the level of standing genetic variation within a species, or the role of selection in the observed divergence. Here, the potential role of selection was analysed by means of  $d_N / d_S$  rates at three levels of divergence: within the Ditrysia, within the more recently diverged *Heliconius* group, and at the intraspecific level by quantifying nucleotide polymorphism within *Pararge aegeria* and *Heliconius erato*. Overall, negative selection was the predominant mode of evolution acting on the PG3 cluster. However, patterns of positive selection were detected operating on the 4 *Shx* genes, both following their duplication and between orthologs in different species. Furthermore, there was evidence for positively selected sites at several positions in the conserved homeodomain, suggesting these differences might be linked to species-specific transcription factor roles. The duplication and divergence of the PG3 cluster in Lepidoptera provides an ideal model system with which to study the rapid evolution of paralogs within the otherwise highly conserved *Hox* cluster.

## II.1. Introduction

*Hox* genes encode transcription factors (TFs) which, amongst other things, specify the identity of segments along the anterior-posterior (AP) axis of segmented bilaterian animals, and are a classical example of the conserved genetic „toolkit“ used during embryogenesis (Pick, 2016; Cheate Jarvela and Pick, 2016; Hrycaj and Wellik, 2016; Pascual-Anaya *et al.*, 2013; Holland, 2013; McGinnis and Krumlauf, 1992). Sequence evolution among these genes is largely constrained, especially within the 60 amino acid long homeodomain motif, which confers target gene specificity through direct DNA binding (Bürglin and Affolter, 2016; Rohs *et al.*, 2009; Joshi *et al.*, 2007; Chu *et al.*, 2012; Gehring *et al.*, 1994). Small sequence changes in *Hox* genes have been shown to result in large body plan differences; for example in *ultrabithorax (ubx)* and the regulation of leg-bearing segments in insects, through the gain of a domain involved in repressing *distal-less (dll)* (Vachon *et al.*, 1992; Estrada and Sánchez-Herrero, 2001; Ronshaugen *et al.*, 2002). Not only are *Hox* gene sequences highly conserved, their chromosomal arrangement has also remained largely unchanged, with the gene order corresponding to their spatio-temporal expression along the animals“ AP axis (Gaunt, 2015; Mann, 1997).

Bilaterian *Hox* clusters are thought to have arisen through tandem duplication of a single ancestral proto-*Hox* gene (Garcia-Fernández, 2005). Following duplication events, theory suggests that one copy maintains its original function, while the other can acquire new functions, through the accumulation of either coding substitutions, or changes in expression patterns (Ohno, 1970; Holland *et al.*, 2017). Paralogs are likely to escape the constraints imposed by deleterious pleiotropic effects which might be caused as a result of direct protein evolution, as the original function can be maintained by one of the copies. *Hox* gene duplication and divergence thus promotes the evolution of segmental identity along the AP axis. *Hox* gene evolution and that of their expression patterns do indeed underpin many spectacular examples of morphological evolution, including changes in segmental identity and the evolution of leg and wing number in insects (Ronshaugen *et al.*, 2002; Averof and Patel, 1997; Galant and Carroll,

2002; Lewis, 1978; Warren *et al.*, 1994). Furthermore, there is evidence that even very small changes in the homeobox can have large consequences. In higher flies, this is demonstrated by the *Hox3*-derived *bicoid* (*bcd*) in which a glycine to lysine substitution at position 50 of the *bcd* homeodomain allowed convergence in functionality on *orthodenticle* (*otd*) as a result (Hanes and Brent, 1989; Hanes and Brent, 1991).

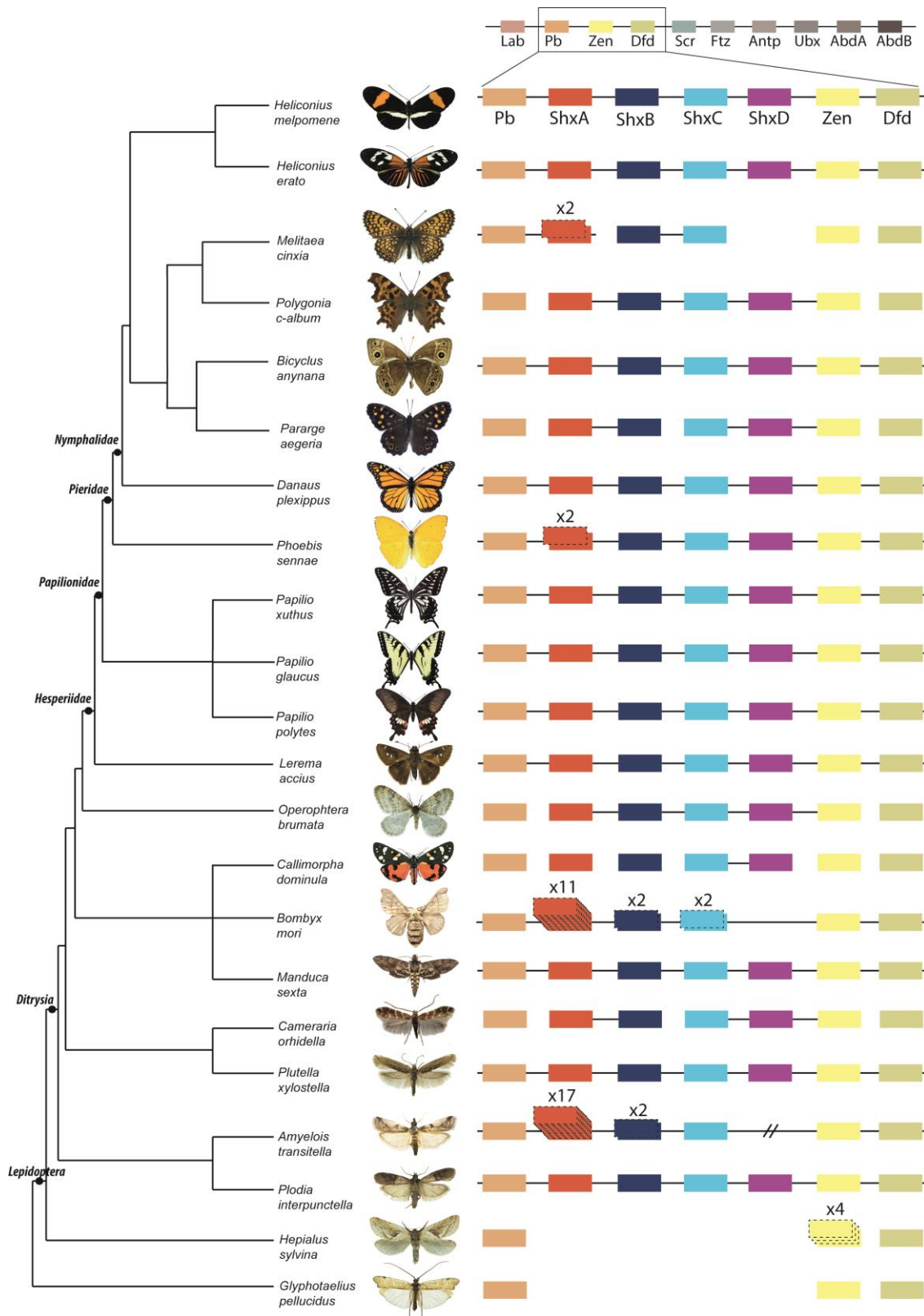
The majority of insect *Hox* genes are also involved in body patterning, with the exception of *Hox3*, whose function has mostly diverged. During the early radiation of the insects, the paralogy group 3 (PG3) gene (i.e. *Hox3*) lost its role in specifying segment identity along the AP axis and attained a role in the specification (and possibly maintenance) of extraembryonic tissue (Hughes *et al.*, 2004; Papillon and Telford, 2007; Schmidt-Ott *et al.*, 2010) (see also Chapter I). Unlike canonical *Hox* genes, the *Hox3* locus is also unusual in that it has also been shown to have duplicated independently in a variety of insects, including the flour beetle *T. castaneum* to yield two copies named *zerknüllt1* (*zen*) and 2 (after the *Drosophila* phenotype) (Falciani *et al.*, 1996; van der Zee *et al.*, 2005), in the hoverfly *E. balteatus* to give seven copies of *zen* (Rafiqi, 2008) and in the fruit fly *D. melanogaster* to give rise to two copies of *zen* and the highly divergent *bcd* (Stauber *et al.*, 1999; Stauber *et al.*, 2002).

More recently, it was shown that several rounds of duplications also occurred during the radiation of the Ditrysia, a derived clade of the Lepidoptera, to give rise to 4 highly divergent *Hox3* genes alongside *zen*, and which were named the Special Homeobox genes (*ShxA-D*; see Figure II-1) (Ferguson *et al.*, 2014). The *Shx* genes likely arose through tandem duplication of the ancestral gene *zen*, and are located in the ditrysonian *Hox* cluster between *proboscipedia* (*pb*) and *Deformed* (*Dfd*). The spatio-temporal expression patterns obtained for the *Shx* genes in the Speckled Wood butterfly *P. aegeria* also strongly suggest a function in specifying extraembryonic tissue – specifically, a region of epithelial cells that covers the embryo during early embryogenesis known as the serosa (Ferguson *et al.*, 2014)(Chapter I). The serosa covers the *P. aegeria* embryo by 12hAEL, by which time zygotic expression of all 4 *Shx* genes becomes localised to the serosal epithelium. In contrast to the canonical *Hox* genes, and the

ancestral *zen* gene, the *Shx* genes display a large amount of sequence variation, both between paralogs as well as between their orthologs in different species (Ferguson *et al.*, 2014 and this Chapter). Furthermore, an excess of non synonymous substitutions was detected following their duplication, suggesting that a period of positive selection may have shaped the patterns of observed variation (Ferguson *et al.*, 2014).

The serosa is implicated in desiccation/drought resistance, processing environmental toxins and mounting an immune response in a range of insects, including beetles and the ditrysian moth *Manduca sexta* (Jacobs *et al.*, 2013; Jacobs *et al.*, 2014; Lamer and Dorn, 2001; Berger-Twelbeck *et al.*, 2003; Orth *et al.*, 2003; Rezende *et al.*, 2008). Experiments performed in *T. castaneum* where serosal-less embryos were exposed to varying levels of humidity (Jacobs *et al.*, 2013), and pathogens (Jacobs *et al.*, 2014), have shown that the serosa is not only required for embryos to survive under low humidity environments, but also to mount an adequate immune response following (invasive) pathogenic exposure. This raises the interesting possibility that selection pressures acting on the ancestral *Hox3* gene have led to its duplication and divergence in species whose ecological niches required a higher resistance to desiccation, toxins and/or infection. The large sequence variation observed for the *Shx* genes in Lepidoptera thus provides the ideal opportunity to further study the patterns of observed variation in relation to possible selection pressures affecting the evolution of an ecologically relevant tissue.

Here, selection acting on the *Hox3* locus in Lepidoptera was examined at three separate levels of divergence. In order to examine large scale sequence variation and associated selection, sequences from the five paralogs were recovered from a wide range of Lepidoptera spanning about 100 to 140 million years of evolution (Wahlberg *et al.*, 2009). Selection acting at the locus was also examined in the more recent *Heliconius* divergence, which spans around 25 million years of evolution (Kozak *et al.*, 2015). Finally, at the micro-evolutionary level, intraspecific sequence variation was assessed, by comparing standing variation within our model species *P. aegeria*, and from recent population genetic data recovered for *H. erato*.



**Figure II - 1. The *Hox3* Locus in Lepidoptera**

Phylogeny of lepidopteran species and their Trichopteran outgroup alongside their respective *Hox3* locus. Linkage along scaffolds is indicated by continuous lines and copy number shown above individual genes. Diagonal lines show poorly defined region of the genome. All sequences recovered from lepbase.org (Chalis *et al.*, 2016).



## **II.2. Materials and Methods**

### **II.2.a Annotation of *Shx* genes in Lepidoptera and comparative analysis between species**

Refinement of *Hox3* annotation across the lepidopteran phylogeny was performed by further annotating the *Shx* genes across species'' genomes found on leabase.org V4.0 (Challis *et al.*, 2016). Annotation was performed by tBLASTn searches of previously identified *Shx* genes against the genomes of 22 Lepidopteran species (see Figure II-1 for species used). Where available, the automatic CDS prediction on leabase.org V4.0 (CDS databases) was used to extract the coding region of the respective paralogs, and checked for linkage along the scaffold (indicated as continuous lines in Figure II-1). Scaffold databases were also searched using other species homeodomains to confirm presence/absence of each gene. Where CDS databases were not available, the scaffold hit was extracted as a FASTA file and manually annotated through alignments of previously identified paralogs to identify intron-exon boundaries and create a CDS annotation using the Ugene platform (Okonechnikov *et al.*, 2012). Such scaffolds containing the putative location of the *Hox3* genes were also verified for linkage to other *Shx* genes and, where available, were annotated in relation to the flanking *pb* and *Dfd* genes. For phylogenetic analyses comparing canonical *Hox* genes to the *Hox3* group, the homeodomain of all *Hox* genes was also annotated and extracted in these species.

### **II.2.b DNA extraction, amplification, sequencing, and alignment in *Pararge aegeria***

In order to study intraspecific variation at the *Hox3* locus and examine possible signatures of selection we sequenced *zen* and the four *Shx* paralogs in ~70 individuals from across Europe (see Chapter III). We further sequenced the homologous regions from the outgroup species *Pararge xiphoides* in order to analyse intraspecific divergence from a recently diverged species (Weingartner *et al.*, 2006).

The five genes were amplified from total genomic DNA extracted from the abdomen of dried adult butterflies using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions. Double-stranded DNA was amplified through polymerase chain reactions (PCR) in 20- $\mu$ L volumes containing: 12.4  $\mu$ L DEPC treated H<sub>2</sub>O, 4 $\mu$ L 5x iProof HF buffer, 0.4  $\mu$ L 10 mM dNTPs, 0.5  $\mu$ L of each primer (10  $\mu$ M), 0.2  $\mu$ L iProof DNA Polymerase (Bio-Rad, 2U/  $\mu$ L) and 1  $\mu$ L of extracted DNA (Appendix III – Note 1 for primers used and cycling conditions). PCR products showing a clear band in gel electrophoresis were purified using QIAquick PCR Purification columns (Qiagen) and sequenced by EurofinsGenomics Ltd.

### **II.2.c Population genetic data from *Heliconius erato* and *Heliconius himera***

*Hox3* sequences for *H.erato* individuals were extracted from previously published genomes (Belleghem *et al.*, 2017). The final dataset contained 104 *H. erato* and 5 *H. himera* individuals. Individual VCF files were searched for respective scaffold locations containing the previously annotated *Hox3* genes in *H. erato* found on Lepbase.org (Challis *et al.*, 2016), and extracted to create individual CDS alignments spanning all individuals. Files were imported, manually checked and aligned in Ugene (Okonechnikov *et al.*, 2012).

### **II.2.d Phylogenetic analyses**

Orthologous sequences were edited and aligned using the MUSCLE algorithm (Edgar, 2004), implemented in Ugene (Okonechnikov *et al.*, 2012) and imported into MEGA 7 software package (Kumar *et al.*, 2016). A best fitting nucleotide substitution model was obtained by using the maximum likelihood model approach in MEGA7. A Maximum likelihood tree was obtained from homeodomain alignment using RAxML and LG+ $\Gamma$  model with 100 bootstrap replicates for analyses spanning all Lepidoptera. The topology inferred from these trees was used in subsequent selection analyses. For large-scale phylogenetic analyses, trees were based on homeodomain sequences only, to ensure accurate alignments. For those involving the *Heliconius* radiation, the entire CDS was used (van Schooten *et al.*, 2016), and an automatic

neighbour joining tree created by the Datamonkey.org (Pond and Frost, 2005) server was used in subsequent analyses.

## II.2.c Selection analyses

Over larger evolutionary time, DNA sequences under pervasive positive selection show an increase in nonsynonymous ( $d_N$ ) over synonymous ( $d_S$ ) substitution rates (Cannarozzi and Schneider, 2012). On the other hand, sequences under purifying or negative selection will show an excess of synonymous substitution rates, as constraints on function penalise nonsynonymous substitutions that alter protein structure. In order to infer selection pressures acting on the *Hox3* paralogs,  $d_N / d_S$  ( $\omega$ ) rates were examined using the package HyPhy (Sergei L. Kosakovsky Pond *et al.*, 2005), implemented in the web based server Datamonkey.org (Pond and Frost, 2005). In order to ascertain selection at the interspecific level, codon alignments based on one sequence per species analysed per gene were used. A combined alignment containing the homeodomain of all paralogs for all species was used to examine evidence of selection along branches of the duplicated genes, by using the branch-site REL (BS-REL) method, which looks for lineages along a phylogeny at which a proportion of sites evolve under  $\omega > 1$  (Pond *et al.*, 2011). The BS-REL algorithm models  $\omega$  as three variables,  $\omega_1$ ,  $\omega_2$  and  $\omega_3$  with the first two remaining between zero and one, and where  $\omega_3$  is  $> 1$ .  $\omega_3$  measures the estimate of positive selection by applying a likelihood ratio test (LRT) to determine if  $\omega_3 > 1$  with a  $p$ -value corrected for multiple testing using the Holms procedure. For the large-scale phylogeny analysis, alignments were restricted to the homeodomains only, as highly divergent sequences between species prevented reliable codon alignments. For analysis at codons outside the homeodomain, the analysis was restricted to the genus *Heliconius*, which shows enough divergence, but not so much as to prevent accurate alignments of the paralogs.

In order to assess selection at the level of codons, selective pressures were measured as the difference between  $\omega$  substitution rates per codon using the fixed-effects likelihood (FEL;

(Kosakovsky Pond *et al.*, 2005)), mixed effects model of evolution (MEME; (Murrell *et al.*, 2012)), random effects likelihood (REL; (Kosakovsky Pond *et al.*, 2005)), Fast Unconstrained Bayesian Approximation (FUBAR; (Murrell *et al.*, 2013)). FEL analyses site by site  $\omega$  without the need for prior distribution assumptions, while REL assumes a prior distribution across sites. Sites under positive selection were also analysed with the FUBAR algorithm, which detects sites under pervasive diversifying selection. We further utilised the MEME algorithm, which allows for the identification of episodic diversification on a subset of branches for a subset of sites, by allowing  $\omega$  to vary across branches (Murrell *et al.*, 2012). This method is particularly useful as most sites might be under negative selection across the phylogeny, while experiencing episodes of positive selection only at specific sites within branches in a phylogeny. For all analyses we used the default option of significance cut-off suggested by the DataMonkey.org server ( $p = 0.1$  for FEL/IFEL, Bayes Factor = 50 for REL,  $p = 0.05$  for MEME and posterior probability = 0.9 for FUBAR).

The single likelihood ancestral counting (SLAC) (Kosakovsky Pond *et al.*, 2005) was subsequently applied to obtain overall  $\omega$  values and to corroborate identification of positively selected sites using other methods. The SLAC method is the most conservative and, therefore, we expect low false positive detection rates. Following the authors' recommendation, we established the level of statistical significance at  $p = 0.1$  (*i.e.*, the default option) for SLAC analyses. For all analyses, an alignment and a corresponding phylogenetic tree are submitted, and the best fitting nucleotide substitution model is automatically selected using the automatic model selection tool on the Datamonkey Server.

For analyses involving allele frequency spectra within *P. aegeria* and *H. erato*, each sequence was duplicated and allocated the respective heterozygous position, using the unphase/Genotype data option implemented in DNAsp. Calculations were carried out over 1,000 iterations, 10 thinning intervals, and 1,000 burn-in iterations. Analyses involving basic polymorphism statistics and departures from neutrality including the McDonald-Kreitman test

(McDonald and Kreitman, 1991), Tajima's D (Tajima, 1989) and Fu and Li's D (Fu and Li, 1993) were performed in DNAsp v.5.0 (Rozas, 2009). We further measured alignment wide  $\Theta$  values using the SLAC method. For intraspecific data, the alignments were first checked for recombination events using the GARD algorithm also implemented in DataMonkey.org.

## II.3. Results

### II.3.a Molecular Evolution of the *Hox3* Locus in Lepidoptera

Following BLAST searches in the leabase database (Challis *et al.*, 2016), we were able to recover mostly complete sequences from 30 new species further to the 9 previously examined (Appendix II – Table 3) (Ferguson *et al.*, 2014). Examination of the *Hox3* locus across the lepidopteran phylogeny revealed further instances of duplication, and gene loss. *ShxA* appears to have duplicated independently in both *Melitaea cinxia* and *Phoebis sennae*, where the duplication is accompanied by a loss of a clear *ShxD* homolog in *M. cinxia*, but not *P. sennae*. Multiple copies of *ShxA* are also present in *B. mori*, as well as a loss of *ShxD* (Chai *et al.*, 2008; Ferguson *et al.*, 2014). Examination of the *Hox3* locus in *Amyelois transitella* recovered 17 *ShxA*-like homeodomains located in tandem, as well as two *ShxB*-like genes. Many of these contain internal stop codons, and multiple homeodomains within a single open reading frame (ORF), suggesting they are likely non-functional. Examination of the *A. transitella* scaffold also failed to recover a *ShxD* homolog. However, the putative location of the gene is in a poorly resolved region of the genome, making it difficult to confirm presence/absence in this species. Annotation of all other species analysed confirmed the retainment of a core set of four *Shx* genes in addition to the ancestral gene *zen*, in accordance with previously published results (Ferguson *et al.*, 2014).

Analysis of the homeodomain within the Lepidoptera revealed a clear excess of synonymous over nonsynonymous substitutions for all *Hox3* paralogs with mean  $\Theta$  values ranging from 0.004 to 0.135 (Appendix II Table - 4). When using a combined alignment

containing the homeodomain of all paralogs to look for evidence of positive selection along branches using the BS-REL method, there was statistical support for  $\omega > 1$  in the branch leading to *ShxB*, *ShxC* and *ShxD* (Node 2 in Appendix II - Figure 1, corrected  $p = 0.005$ ), in accordance with previous results (Ferguson *et al.*, 2014). This may suggest that following the duplication of *zen*, positive selection acted on a subset of sites within the homeodomain of the three paralogs (Table II-1). Evidence of  $\omega > 1$  on node 32 was also found (Table II-1), corresponding to the internal branch leading to *ShxC* (corrected  $p = 0.018$ ).

**Table II - 1. Positive selection following gene duplication at the *Hox3* locus in *Ditrysia* as determined using the HyPhy branch-site random effects model (Pond *et al.*, 2011).**

Branch	Mean $\Omega$	$\Omega 1$	p1	$\Omega 2$	p2	$\Omega 3$	p3	LRT	Holms $p$ -value
<i>ShxB*ShxC*ShxD</i>	0.969	0	0.712	0	0.001	10000	0.287	16.024	0.005
<i>ShxC</i>	0.418	0	0.823	0	0.021	3333	0.156	13.413	0.018

When applying methods to look for specific sites under selection across the combined *Hox3* alignment, none are found. It is possible that there is not sufficient signal to detect individual sites under selection, but that BS-REL recovers enough power when the signal is pooled among sites.

When looking for specific codons under selection between orthologs using MEME, we find evidence of episodic diversifying selection on positions 37 and 54 of the homeodomain of *ShxB* ( $p=0.04$  and  $0.05$  respectively), at position 1 of *ShxC* ( $p=0.05$ ) and at position 43 of *ShxD* ( $p=0.05$ ) (Appendix II - Figure 2). Other methods also show an excess of nonsynonymous substitutions at these sites but no statistical support. Interestingly, position 54 of the homeodomain is within the third alpha helix, a position known to make specific contact with DNA (Mann *et al.*, 2009). Furthermore, while not significant for positive selection, it is interesting to note that all *Papilio* species analysed have a fixed K50 substitution within the highly conserved WFQNRR motif in the third alpha helix of the homeodomain. This is a

position known to make specific DNA contact, and is argued to be a key substitution following the divergence of *bcd* in cyclorrhaphan flies (Hanes and Brent, 1991; Hanes and Brent, 1989).

When restricting the analysis to full length alignments of the *Hox3* genes within the *Heliconius* group, we also observed an overall excess of synonymous over nonsynonymous substitutions (Table II - 2). Overall  $\omega$  values were higher than those observed for alignments restricted to the Homeodomain, but followed the same overall trend with *zen* displaying highest overall purifying selection, followed by *ShxA*, *ShxC*, *ShxD* and *ShxB*. In total, there is evidence of 79 positively selected sites across the 5 paralogs using at least one method. Notably, during the *Heliconius* radiation, MEME found evidence of positive diversifying selection on position 23 of the homeodomain for *ShxA*, within the first alpha helix (Table II – 3; Appendix II - Figure 3). Both SLAC and FEL also find significant evidence for positive selection at this site. Evidence of positive selection was also found acting on position 18 of the homeodomain of *ShxB* using three methods (FEL, IFEL and MEME) - also within the first alpha helix. Position 19 of the homeodomain of *ShxB* was also significant for episodic selection. MEME found further evidence of selection on position 36 of the homeodomain of *ShxB*, within the second alpha helix. For *ShxD*, we found evidence of positive selection at position 33 of the homeodomain, within the second alpha helix (Table II – 3.). See Appendix II - Table 5 for results on all sites under selection across the 5 genes.

**Table II - 2 Summary of Selection acting on the *Hox3* genes in *Heliconius* butterflies**

<i>Heliconius</i> Radiation									
Alignments			Sites under selection						
Gene	Taxa	bp (aa)	$\omega$ (95% CI)	SLAC	FEL	IFEL	REL	MEME	FUBAR
<i>ShxA</i>	20	1080 (360)	0.31 (0.28, 0.33)	1	8	7	0	15	0
<i>ShxB</i>	19	1056 (352)	0.41 (0.38, 0.44)	1	8	11	3	16	1
<i>ShxC</i>	19	756 (252)	0.28 (0.25, 0.31)	0	4	3	0	12	1
<i>ShxD</i>	16	855 (285)	0.37 (0.33, 0.41)	0	1	2	0	9	0
<i>zen</i>	20	1665 (555)	0.15 (0.13, 0.16)	3	5	3	5	9	2

**Table II - 3 Sites across the *Heliconius* homeodomains under positive diversifying selection as identified by MEME.**

MEME <i>Heliconius</i> Homeodomains					
Gene	HD position	$\alpha$	$\beta$ -	Pr[ $\beta=\beta$ -]	<i>p</i> -value
<i>ShxA</i>	23	0.00	0.00	0.77	0.01
<i>ShxB</i>	18	0.00	0.00	0.73	0.03
	19	1.03	0.74	0.92	0.05
	36	0.00	0.00	0.94	0.02
<i>ShxD</i>	33	0.93	0.70	0.87	0.05

### II.3.b Molecular Evolution of the *Hox3* locus within species

In order to study intraspecific variation at the *Hox3* locus in *P. aegeria* and *H. erato*, we examined DNA sequence variation in the coding region of the four *Shx* paralogs as well as the ancestral *zen* gene. Since we were interested in obtaining as much intraspecific variation as possible, we sampled individuals from a range of populations from geographically widespread locations (see also Chapter III for a detailed discussion on the biogeography and relatedness of these populations in *P. aegeria*, and Belleghem, 2017 for *H. erato* populations). For *P. aegeria*,



we obtained a total of ~70 sequences per gene, from a total of 35 locations throughout Europe (See Chapter III for details). We further sequenced the 5 *Hox3* genes in the outgroup species *P. xiphiodes*, in order to compare standing variation in *P. aegeria* to a recently diverged species. For *H. erato*, we obtained 104 sequences per gene, and a further 5 sequences from the closely related *H. himera*.

Various tests based on the frequency spectra of the *Hox3* locus were performed to assess levels of polymorphisms in both species (Table II - 4 and Table II – 5). For *P. aegeria*, average pairwise nucleotide diversity ( $\pi$ ) was similar between paralogs, except for *ShxD*, which showed a larger level of nucleotide variation. All paralogs under investigation showed a trend towards a negative Tajima's D value (except for *ShxD* (0.364)) indicating a deficit of high frequency segregating sites. However, none were significant. When applying Fu and Li's D\* statistic, using *P. xiphiodes* as an outgroup to polarise mutations, we obtained significant negative values for *ShxB*, *ShxC* and *zen*, indicating a possible excess of low-frequency variants in these genes. This could be indicative of a population expansion following a bottleneck (Chapter III), or a selective sweep, either because of linkage to a neighbouring gene, or an effect of the locus itself. Average nucleotide differences between *P. aegeria* and *P. xiphiodes* were also similar, except for *zen*, which showed a lower level of differentiation between species. Furthermore, the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) was performed, which measures the ratios of synonymous to nonsynonymous substitutions within and between species, using *P. xiphiodes* as an outgroup. The MK tests show positive Neutrality Index (NI) values for the all paralogs analysed, which are indicative of negative selection. However, only *zen* was significant (NI=4.375,  $p=0.02$ ). This is indicative of negative selection acting on the locus.

Similar patterns were observed for *H. erato*. Average pairwise nucleotide diversity was also similar between paralogs, with *ShxD* also being the most variable. Negative Tajima's D was also recorded, and was significant for *ShxB*, *ShxC* and *zen*. Using *H. himera* as an outgroup,

Fu and Li's D was further significant for *ShxA*. The MK test could not be performed in this species, as no nonsynonymous divergence was observed between *H. himera* and *H. erato*.

Interestingly, several replacement polymorphisms were observed within the homeodomains of several genes in both *P. aegeria* and *H. erato* (Appendix II - Tables 1 and 2). A number of these polymorphisms occur at relatively high frequencies, and at positions known to make specific DNA contact within the third alpha helix of the homeodomain. Of these, *ShxB* was the most polymorphic, with 9 non-synonymous changes occurring at frequencies of >10 within the populations analysed.

**Table II – 4. Summary statistics and selection tests at each of the *Hox3* paralogs in *Pararge aegeria*<sup>1</sup>.**

Gene	bp (aa)	SNPs(syn:nsyn)	$\omega$	$\theta_w$	$\pi$	D	d	MK	$K_{aeg-xip}$
<i>ShxA</i>	717(239)	20 (9:11)	0.19	0.005	0.003	-0.972	-0.972	1.222	0.036
<i>ShxB</i>	1077(359)	61 (21:40)	0.51	0.010	0.006	-1.331	-3.559**	0.784	0.028
<i>ShxC</i>	645(215)	18 (6:12)	0.36	0.009	0.006	-1.026	-2.883*	1.267	0.038
<i>ShxD</i>	477(159)	26 (11:15)	0.26	0.013	0.014	0.364	-0.689	1.875	0.031
<i>zen</i>	1599(533)	45 (20:25)	0.39	0.006	0.005	-0.527	-2.774*	4.375**	0.016

<sup>1</sup>Alignment size, expected mutation rate ( $\theta_w$ ) and average pairwise differences ( $\pi$ ) are shown. The selection test results for Tajima's D (D), Fu and Li's D (d), McDonald-Kreitman (MK), Kaeg-xip showing nucleotide differences between *P. aegeria* and *P. xiphiodes* and associated statistical significance are shown (\*: p= 0.05; \*\*: p= 0.01). The counts of synonymous to nonsynonymous SNPs are also shown (SNPs) as well as overall  $\omega$  values with 95% confidence intervals as estimated by the SLAC method.

**Table II – 5. Summary statistics and selection tests at each of the *Hox3* paralogs in *Heliconius erato*.<sup>1</sup>**

Gene	bp(aa)	SNPs(syn:nsyn)	$\omega$	$\theta_w$	$\pi$	D	d	K <sub>era-him</sub>
<i>ShxA</i>	1179(393)	246(147:99)	0.19	0.0433	0.0245	-1.590	-5.085**	0.024
<i>ShxB</i>	1014(338)	121(46:75)	0.47	0.0359	0.0082	-2.429**	-5.507**	0.013
<i>ShxC</i>	738(246)	77(42:35)	0.28	0.0313	0.0096	-2.101**	-0.187	0.010
<i>ShxD</i>	804(268)	139(69:70)	0.40	0.0491	0.0271	-1.583	-0.633	0.030
<i>zen</i>	1653(551)	341(216:125)	0.16	0.0381	0.0174	-1.844*	-4.969**	0.017

<sup>1</sup>Alignment size, expected mutation rate ( $\theta_w$ ) and average pairwise differences ( $\pi$ ) are shown. The selection test results for Tajima's D (D), Fu and Li's D (d), Kaeg-xip showing nucleotide differences between *H. erato* and *H. hamera* and associated statistical significance are shown (\*:  $p=0.05$ ; \*\*:  $p=0.01$ ). The counts of synonymous to nonsynonymous SNPs are also shown (SNPs) as well as overall  $\omega$  values with 95% confidence intervals as estimated by the SLAC method.

## II.4. Discussion

This study analysed the potential selective pressures operating on a duplicate set of *Hox3* genes at three separate levels of divergence; within the Ditryisia, within the more recently diverged *Heliconius* group, and at the intraspecific level by quantifying nucleotide polymorphisms within *P. aegeria* and *H. erato*. As expected from a set of highly conserved genes, overall, there was evidence for a conservative mode of evolution, where negative selection is the predominant force acting on the *Hox3* genes. However, significant instances of positive selection following both the duplication of the *Shx* genes and between paralogs were found, suggesting periods of positive selection affected the duplicates and individual residues between orthologs. Furthermore, several of the positively selected sites were found within the homeodomain, suggesting that following both the duplication of *Hox3* in Ditryisia, and during the divergence between paralogs, the residues affected could have evolved to recognise different target sequences and/or have acquired new residues capable of producing novel protein-protein interactions, leading to a divergence in function between the paralogs as well as between orthologous sequences in different species. An *in silico* evolution model approach limited to the

homeodomain does indeed suggest that the different paralogs are able to recognise separate target motifs (Ferguson *et al.*, 2014), suggesting that the paralogs have diverged in function during Ditrysian evolution.

Interestingly, a relatively high amount of nonsynonymous variation was also observed between the more recently diverged *Heliconius* species, where positively selected sites were also detected within the homeodomain of three of the paralogs. This suggests that the *Shx* genes are under strong selective pressure, even over relatively shorter evolutionary periods. It is also striking to note the high level of nonsynonymous polymorphisms observed segregating through both the *P. aegeria* and *H. erato* populations, some of which segregate at relatively high frequencies even within the homeodomain. These patterns of large divergence might in part be explained by a possible redundancy in function of the paralogs. Slightly deleterious changes segregating through populations and between species might be tolerated if other paralogs were able to compensate for these differences in function. It is therefore possible that due to this compensatory nature, the *Shx* genes might be evolving under relaxed constraints on sequence evolution.

While this study does not provide any direct evidence for the functional significance of the variation observed, evidence of positively selected sites within the homeodomain is interesting for several reasons. Firstly, homeodomains are amongst the most conserved protein motifs described, suggesting any sequence variation is heavily penalised in terms of fitness (i.e. under strong negative selection). The extent of sequence variation between orthologs is thus very unusual for homeobox class genes. Second, there is evidence to suggest that even small changes at the level of the homeodomain can have large implications for target recognition. For example, changing a single amino acid at the C-terminus of the third alpha helix of the protein Paired, can change its binding specificity to that of either *ftz* or *bcd* (Treisman *et al.*, 1989). Similarly, a glycine to lysine substitution at position 50 of the *bcd* homeodomain alters recognition of DNA binding motifs to TAATCC (recognized by K50 class *Hox* proteins such as Otd) from the TAAT(T/G)(A/G) motif recognised by the Antennapedia and Engrailed classes

(Hanes and Brent, 1989). Interestingly, the three *Papilio* species for which *ShxD* sequences are available all show the same K50 substitution at this position, suggesting a possible alteration in binding site specificity as compared to other species. Furthermore, there is evidence for positively selected sites at positions known to make specific DNA contact within other *Shx* paralogs, indicating that these changes are likely non-neutral and of potential adaptive significance. Positively selected sites were also found at several positions outside the homeodomain during the *Heliconius* radiation. While these are not found in any known described functional domains, they could have potentially evolved to affect the overall 3D protein structure and/or have produced novel protein-protein interaction domains. Mutations outside the conserved homeodomain of *Hox* genes have indeed been linked to the evolution of insect body plans (Ronshaugen *et al.*, 2002; Galant and Carroll, 2002). Loss of abdominal appendages in insects is in part explained by coding changes in the Ubx protein, whereby the acquisition of a novel domain allowed it to repress *distal-less (Dll)*. *Dll* is a key gene involved in the development of appendages (Cohen *et al.*, 1989), and its repression in abdominal segments by Ubx coincides with the evolution of the limb-less abdomen in insects (Galant and Carroll, 2002; Grenier and Carroll, 2000).

The evolution of *ftz* is also interesting in this respect, and draws some parallels to *Hox3* evolution in insects. Like *Hox3*, *ftz* is found within the insect *Hox* cluster, but is also not involved in specifying segment identity along the AP axis in *Drosophila*, and instead functions as a pair-rule segmentation protein (Wakimoto *et al.*, 1984; Pick, 2016; Löhr *et al.*, 2001). The loss of function as a canonical *Hox* gene and acquisition of pair-rule function coincides in part with the evolution of a novel motif which is required for the interaction with the novel cofactor *ftz-fl*, and with the loss of the YPWM motif, which mediates *Hox* protein interaction with *Hox* cofactor *extradenticle (Exd)* (Löhr and Pick, 2005). The gain and loss of these motifs have been argued to be key changes that allowed the diversification in function of *ftz* (Heffer *et al.*, 2013).

Several potentially interesting patterns of gene loss and further rounds of duplications were also reported within the Lepidopteran phylogeny. Previous studies had shown that *B. mori*

contained multiple copies of each *Shx* gene, while lacking a *ShxD* homolog (Ferguson *et al.*, 2014; Chai *et al.*, 2008). Due to previous phylogenetic sampling, this was considered an oddity, as many of the copies contained truncated homeodomains and are likely non-functional, and all other species analysed retained a set of four core *Shx* genes. However, the further sampling in this study suggests a further round of independent duplication of *ShxA* in *A. transitella*, *M. cinxia* and *P. sennae*. The duplication of *ShxA* in *M. cinxia* is accompanied by an apparent loss of a *ShxD* homolog, while being retained in *P. sennae*. While scaffold linkage across the *M. cinxia* genome is not well resolved for the locus, a manual annotation of transcript reads over the *Hox3* locus appears to confirm the loss of *ShxD* in *M. cinxia* (Ahola *et al.*, 2014, Sup. Note 8). Searches for conserved *ShxD* motifs and ORFs in the region failed to recover an *ShxD* homeodomain, but showed that limited regions of sequence similarity with other nymphalid *ShxD* sequences were present, albeit highly degraded and with multiple stop codons at each ORF. This suggests that *ShxD* has become highly degraded in *M. cinxia*. The duplication of *ShxA* in *A. transitella*, like *B. mori*, presents an unusual case of multiple duplications. Many of the homeodomains are truncated, and contain other likely deleterious mutations, suggesting they are likely non-functional and perhaps not transcribed. Unfortunately, no transcriptomic data for early developmental stages are available for this species, so it is hard to conclude whether these genes are still being transcribed or have become pseudogenised.

The patterns of duplications and loss could also be explained through the action of gene conversion. This is the phenomenon by which a homologous DNA fragment is transferred to the corresponding location in another paralogous region, leading to identical loci (Chen *et al.*, 2007). The location of the *ShxA* paralogs in these species argues against this. If a gene conversion event had occurred between *ShxD* and *ShxA* in these species, the position of the converted gene along the chromosome should be maintained, leading to a *ShxA* paralog being located adjacent to *zen*. However, linkage along the *P. sennae*, *B. mori* and *A. transitella* scaffolds shows all *ShxA* paralogs are located in tandem along the chromosome, arguing against a gene conversion event.

The arthropod body plan has evolved under a largely constrained set of *Hox* genes. The conservation in both number and sequence of *Hox* genes has led to the widely accepted idea that *Hox* proteins tend to not significantly diverge in function. The evolutionary history of the *Hox3* locus represents one of the few recorded cases of tandem gene duplication within a *Hox* cluster, followed by extensive sequence divergence and instances of both sub- and neo-functionalisation (Schmidt-Ott *et al.*, 2010). Its loss of function as a canonical *Hox* gene and redeployment in specifying extraembryonic tissue, as well as species-specific redeployment in early AP specification in *Drosophila* through the action of *bcd*, demonstrates that some of the most conserved proteins described can diverge and acquire new functions during evolution. Furthermore, duplication and divergence of *Hox3* does not seem to be limited to the cyclorraphan flies, and has occurred independently in several insect lineages, with evidence of neo-functionalisation also observed in *T. castaneum*, where *zen-1* is involved in specifying the serosa, and *zen-2* is mostly involved in dorsal closure (van der Zee *et al.*, 2005). In the hoverfly *E. balteatus*, seven *zen*-like genes have been reported at the *Hox3* locus, where expression patterns differ between the paralogs, also suggesting possible neo and/or subfunctionalisation following their duplication (Rafiqi, 2008). Finally, it appears that radical and significant changes to *Hox3* class homeodomain proteins in Lepidoptera have also occurred, and have likely acquired new and specific roles during Lepidopteran evolution.

Why then is the *Hox3* locus in insects so amenable to duplication? Tandem gene duplication is thought to arise through unequal cross-over events, where cross-over occurs at non-homologous points (Hastings *et al.*, 2009; Reams and Roth, 2015). Chromosomal positioning and sequence repeats can promote these cross-over events (Levinson and Gutman, 1987), resulting in genomic hotspots for gene duplication. It is possible that the chromosomal location and/or potential repeat regions around the *Hox3* locus in insects is characterised by such hotspots, promoting the amenability of the locus to gene duplication. A further reason may lie in its transition from a canonical *Hox* gene to its role in specifying extra embryonic tissue. The role of *Hox* genes in specifying segments along the AP axis is a crucial developmental process, and any dosage effect caused by duplications is likely to be deleterious to the

developing organism. The transition to extraembryonic specification likely relaxed some of the constraints on the locus, making it more amenable to duplication and divergence. Furthermore, the conserved co-linearity of the *Hox* cluster suggests that positioning along the chromosome is crucial to their function, due to either shared enhancers and/or chromatin organisation at the locus (Gaunt, 2015). Expression of the *Hox3* locus likely relies on separate enhancers which direct its expression in extraembryonic tissue, and cross talk to other *Hox* genes is probably limited, thus freeing it from evolutionary constraints operating from other *Hox* genes. Tandem duplication of *zen* is therefore less likely to disrupt any regulatory logic within the *Hox* cluster, and the functional redundancy following the initial duplication can then allow for mutations at both coding and regulatory regions, leading to modification in protein function and expression patterns.

Moreover, the development of an extraembryonic epithelium in insects coincides with their transition to terrestrial habitats (Jacobs *et al.*, 2013; Zeh *et al.*, 1989), and has been shown to be directly involved in protecting the developing embryo against desiccation and pathogen exposure (Jacobs *et al.*, 2013; Jacobs *et al.*, 2014; Lamer and Dorn, 2001; Rezende *et al.*, 2008). Selective pressures could therefore be acting at the *Hox3* locus in species whose environmental conditions require the ability to mount a strong immune response, and/or protect against desiccation. In this respect, maternal contribution of Shx proteins in butterflies seems to be a crucial investment in specifying extraembryonic tissue early in development, where *ShxC* already prefigures the future position of the serosa in the developing oocytes (Ferguson *et al.*, 2014). This expression pattern is amongst some of the most complex described, and seems to suggest that early specification of extraembryonic tissue is of great importance to Lepidopteran embryology. The serosa in Lepidoptera is also unusual as it seems to remain intact during the entire duration of embryonic development (Braak *et al.*, in prep, Kobayashi *et al.*, 2003), whereas in most other insect species it is reabsorbed during earlier stages. This suggests that the specification and maintenance of the serosa is likely a key adaptation for successful embryonic development in Lepidoptera.



Overall, the duplication and divergence of the *Shx* genes in Lepidoptera provides an ideal model with which to study the fate of duplicate genes in the context of an ecologically relevant tissue. Many questions still remain with regards to the roles the *Shx* genes play in Lepidopteran biology. Most importantly, direct experimental evidence is required to show that the positively selected sites found in this study are of functional significance. Are the substitutions found between paralogs able to confer different target specificity between the duplicates? And furthermore, are differences between orthologous sequences adaptive? While no direct evidence is yet available to answer these questions, the fact that positive Darwinian evolution has acted on key residues of the homeodomains hints that this is likely to be the case.

## Chapter III

*Pararge aegeria* displays patterns of geographic variation in the novel *Shx* genes distinct from its biogeography as inferred by *COI*

## Abstract

Speckled Wood butterflies (*Pararge aegeria*) are an emerging model system in ecology, phylogeography and developmental studies, but so far detailed phylogeographical data is lacking for this species. *P. aegeria* have a widespread distribution, experiencing a large variety of environmental conditions. Such species have the potential to reveal broad biogeographical patterns, thus providing valuable data on how species react to environmental pressures across large geographic ranges. In particular, understanding the way species respond to past and current climatic condition is of major interest as these can have large effects on a species ability to adapt to changing environments. Here, I studied the geographic distribution of *P. aegeria* by sequencing the mitochondrial gene *cytochrome oxidase subunit I (COI)* and the nuclear genes *wingless* and the five *Hox3* genes. Analysis of mitochondrial sequences revealed the presence of two main lineages separated by the Mediterranean. Distinct *COI* lineages are maintained over relatively short distances, such as between the islands of Sardinia and Corsica. However, mating experiments revealed no evidence of reproductive isolation between the lineages, and nuclear genes suggest gene flow between sea straits is possible. I propose that following the post-glacial recolonisation of Europe, the ancestral *COI* lineage was maintained in N. Africa and Mediterranean islands, while a new *COI* variant colonised from E. Europe, replacing and outcompeting the ancestral variant. Several hypotheses will be discussed that may explain the discordance between *Shx* genes, *COI* and *wg*, including male-biased dispersal, selection and differential rates of gene evolution.

### III.1. Introduction

As discussed in Chapter II, the four *Shx* genes (*ShxA*, *ShxB*, *ShxC*, and *ShxD*) are unique duplications of *Hox3* in the Ditrysia, of which the ancestral gene is called *zerknüllt* (*zen*) in insects. These genes show a tremendous pattern of genetic variability, both between species across and within genera, and even within a species, as demonstrated using the model system the Speckled Wood butterfly, *Pararge aegeria*. The polymorphisms, and the signatures of selection, identified in Chapter II are intriguing as we do not fully understand the selection pressures operating on these genes. Although they play a key role in the specification and possibly the maintenance and functioning of the serosa, these genes do have pleiotropic effects, evidenced by the later embryonic (as opposed to extraembryonic) expression patterns (Chapter IV). In this Chapter I take a different approach to investigate the evolution of the *Shx* genes, by examining the patterns of geographical variation in the polymorphisms identified in Chapter II. These patterns are then to be compared against the biogeography of *P. aegeria*. In order to do so accurately, this chapter will also significantly update what we know so far about the biogeography of *P. aegeria* (Weingartner *et al.* 2006; Habel *et al.* 2013; Tison *et al.* 2014), in particular with respect to populations for which we have sequenced the *Shx* genes.

Climatic fluctuations associated with the Pleistocene period in Europe have had a tremendous impact on the emergence of different lineages for many temperate species (Cooper *et al.* 1995; Taberlet *et al.* 1998; Seddon *et al.* 2001). During cold periods most European species were presumably restricted to Mediterranean areas. Due to the geographic configuration of the Mediterranean region, a series of areas, separated by mountain chains and sea channels have been hypothesised to function as differentiation centres and observation on the genetic variation on many organisms provided wide evidence for this hypothesis (reviewed in Hewitt 1999; Hewitt 2000; Schmitt 2007). Typically, these areas have been identified in the Iberian and Italian Peninsulas and the Balkans for Europe, which have been isolated from each other to various degrees during the long cold periods that characterized the Pleistocene. The large Mediterranean islands, Maghreb and Asia Minor represented further important refugia and

centres of differentiation for species living in the Mediterranean area (Habel *et al.* 2009; Habel *et al.* 2011; Dapporto *et al.* 2011, 2012).

Following the long isolation, many species evolved different lineages which, in some cases, have been proposed to represent recently evolved sister species (e.g. Dapporto, 2009; Ribera and Volger, 2004). During the relatively short warm periods, thermophilic species that were constrained to these areas began northwards expansions and re-colonised previously glaciated habitats. There is a pervasive signal in patterns of post glacial expansion showing that although lineages and sister species can expand over thousands of kilometres in Europe, when they meet in re-colonized areas they tend to form only very narrow areas of overlap or they establish contact zones along even short sea straits (Waters 2011 for a review, Dapporto *et al.* 2011, 2012; Vodă *et al.* 2015a,b; Habel *et al.* 2017 for Mediterranean butterflies). Several explanations have been proposed to explain these distribution patterns and they involve density dependent phenomena, climatic and environmental preferences, reproductive interference, dispersal limitations and/or competitive exclusion (Waters 2011; Vodă *et al.* 2015b, 2016). Due to the high number of potential mechanisms that can concur to determine patterns of mutual exclusion, understanding the processes responsible for the observed distributions requires highly multidisciplinary approaches (Vodă *et al.* 2015b for Mediterranean butterflies). Studying the spatial distribution of highly diverging genetic lineages and their tendency to form more or less extended parapatric areas of contact, has fundamental implications in understanding the onset of the speciation process (e.g. Arias *et al.* 2008, Habel *et al.* 2017 for butterflies in particular).

The Speckled Wood butterfly, *Pararge aegeria* represents a potentially ideal model to study the distribution of genetic lineages and their spatial segregation. In fact, it has a widespread distribution (ranging from the Maghreb, throughout Europe and reaching western Asia), experiencing a large variety of environmental settings, from cold and wet conditions in northern Europe to hot and dry conditions in southern Europe and North Africa (Weingartner *et al.* 2006; Habel *et al.* 2013; Tison *et al.* 2014). Moreover, this species occurs in many Mediterranean and Atlantic islands, thus allowing the study of dispersal both over ground and

across sea straits. Such widely distributed species have the potential to reveal broad biogeographical patterns associated with responses to both biotic and abiotic factors and to the evolution of different lineages, thus providing valuable data on how species react in time and space to environmental pressures across large geographic areas (Parmesan 1999; Oliver *et al.* 2015).

Based on the variation in the mitochondrial cytochrome *c* oxidase subunit I (*COI*) gene, in the nuclear *wingless* gene and in microsatellites two main lineages of *P. aegeria* have been identified (Weingartner *et al.* 2006; Habel *et al.* 2013; Dapporto *et al.* 2017). The first occurs in Maghreb, in Balearic Islands and in Sardinia, the second in mainland Europe and Asia. Accordingly, the two lineages are separated by three sea channels, the Gibraltar strait between Morocco and Spain, the Sicilian channel between Sicily and Tunisia and the Bonifacio channel between Sardinia and Corsica (Dinca *et al.* 2015; Voda *et al.* 2016; Dapporto *et al.*, 2017). The differentiation between Corsican and Sardinian populations of *P. aegeria* is also evident at the morphological level with a divergence in male genital shape between the two lineages (Dapporto *et al.* 2012). Furthermore, a recent study by Longdon *et al.* (2017) examining the population genetics (which often reflect those of their hosts, Wilfert & Jiggins 2014; Longdon *et al.* 2017), and modes of transmission in a range of different Rhabdoviruses, highlighted discrete Sardinian and Corsican populations of the *P. aegeria* specific Rhabdovirus PAegRV, suggesting limited dispersal between the islands (for a detailed description of this recently discovered virus see Longdon *et al.* 2015).

The variation between populations on Corsica and Sardinia represents a particularly intriguing case. Indeed, these islands are separated by less than 11 km of sea straits and several small adjacent islands can act as stepping stones. Moreover, these islands were connected during the last glacial maximum and this suggests that the two different populations have been established from different source populations following relatively recent post glacial dynamics (Dapporto 2010) and thereafter there has been little or no dispersal over the Bonifacio strait.

Several explanations can be provided for the observed distributions of island populations. Corsica and Sardinia have different environmental settings, with considerable

variation in temperature and rainfall (reflected in the vegetation) (Hijmans *et al.* 2005; Zinetti *et al.* 2013). It is highly unlikely that climatic differences alone prevent the European lineage from establishing populations on Sardinia and vice versa, however local adaptation may reduce the likelihood of colonisation (cf. Richter-Boix *et al.* 2013). Climatic factors, and their effects on host plants, have indeed been shown to have strong selection pressures in *P. aegeria* that could result in different egg-laying strategies (Hill *et al.* 1999a; Hill *et al.* 1999b). Furthermore, it may be possible that reproductive isolation is emerging between the two lineages preventing migrants from establishing in local populations on the other island. Female mate choice, in particular, has been recorded as a factor in maintaining reproductive isolation in several butterfly species (e.g. Friberg *et al.* 2008; Dincă *et al.* 2013; Pinzari & Sbordoni 2013).

Even in the absence of reproductive barriers, hybrid fitness could be reduced, thus explaining the mutual exclusion pattern. Although very little is understood about the reduction in hybrid fitness at the molecular level (Barton & Hewitt 1985; Presgraves *et al.* 2003; Rogers & Bernatchez 2006), three specific forms of post-zygotic isolation have been described: sterility of F1 hybrids, lethality of F1 hybrids and degeneracy of F2 hybrids (Dobzhansky 1970; Dumas *et al.* 2015). Thus, it may very well be possible that no strict pre- or postzygotic barriers exist, but that immigrants and their (hybrid) offspring find themselves at a selective disadvantage compared to local populations.

In order to refine the known biogeographical information on *P. aegeria*, and to take into account the aforementioned issues, we sampled numerous populations of *P. aegeria* across Europe and North Africa, with a special focus on Corsica and Sardinia. *COI* from 346 individuals was sequenced, as well as the nuclear developmental gene *wingless* (*wg*) for a subset of individuals spanning this range. Furthermore, in order to examine the geographical variation in the genetic polymorphisms identified in Chapter II, against the *P. aegeria* biogeography, we sequenced the five *Hox3* genes (for a description see Ferguson *et al.* 2014 and other Chapters), for a subset of 70 individuals. This enabled the investigation of the distribution of the two genetic lineages and their intra-lineage genetic diversity over the study area to a high spatial

resolution. Moreover, to test the hypothesis that pre- or postzygotic barriers limit gene flow between the two lineages over Sardinia and Corsica the reproductive strategies of Sardinian and Corsican *P. aegeria* females were examined by mean of courtship and mating experiments.

## **III.2. Material and Methods**

### **III.2.a Study species**

Between the 6<sup>th</sup> and the 12<sup>th</sup> of May 2014 *P. aegeria* females were collected in the field from 11 different localities in: Sardinia (Aritzo, Desulo and Tempio Pausania), Corsica (Asco, Zonza, Bavella, Bonifacio, Solenzara, Cavallo Morto and Pietralba) and La Maddalena (Sualeddu), a smaller island off the north coast of Sardinia. In total 32 females were caught, 18 from Corsica, 13 from Sardinia and 1 from La Maddalena. Eggs from collected females were obtained *in-situ* and brought to the laboratory in Oxford, UK. Upon hatching, larvae from these eggs were reared on a mix of host plants known to be used by *P. aegeria* in Europe (a mix of *Poa trivialis* and *Dactylis glomerata*) and reared at  $21 \pm 2^\circ\text{C}$  (60% RH, 16L:8D) (cf. Breuker *et al.* 2007; Gibbs *et al.* 2010b). The females collected in the field laid readily on these plant species, as did all the females used in this experiment (cf. Breuker *et al.* 2007; Gibbs *et al.* 2010b). Pupae were sexed and kept individually, to ensure virgin adults were available for setting up crosses. A total of 22 of the 32 field-collected females provided a sufficient number of adult offspring butterflies to perform the crosses; collection details from these females are shown in Table III - 1.



**Table III - 1. Collection details of the 22 females caught in the field and whose offspring were used in the laboratory crosses.**

<b>Female number</b>	<b>Date of collection</b>	<b>Locality</b>	<b>Island</b>	<b>Latitude</b>	<b>Longitude</b>	<b>Altitude</b>
1	06/05/2014	Aritzo	Sardinia	39.9649722	9.13986111	641.32031
2	06/05/2014	Aritzo	Sardinia	39.9454166	9.19938888	1003.49414
3	06/05/2014	Aritzo	Sardinia	39.9454166	9.19938888	1003.49414
6	06/05/2014	Aritzo	Sardinia	39.9454166	9.19938888	1003.49414
7	06/05/2014	Aritzo	Sardinia	39.9454166	9.19938888	1003.49414
8	08/05/2014	Sualeddu	La Maddalena	41.2335	9.40855555	94.334473
9	06/05/2014	Aritzo	Sardinia	39.9649722	9.13986111	641.32031
11	08/05/2014	Tempio Pausania	Sardinia	40.911428	9.095736	480
14	10/05/2014	Zonza	Corsica	41.7521111	9.19152777	792.48657
16	10/05/2014	Tempio Pausania	Sardinia	40.911428	9.0957	480
17	10/05/2014	Bavella	Corsica	41.8106666	9.24669444	698.75878
18	10/05/2014	Zonza	Corsica	41.7594722	9.18419444	848.48266
20	09/05/2014	Bonifacio	Corsica	41.3772222	9.17944444	83
22	10/05/2014	Zonza	Corsica	41.7521111	9.19152777	792.48657
23	07/05/2014	Desulo	Sardinia	40.0375	9.25694444	975
24	10/05/2014	Bavella	Corsica	41.8106666	9.24669444	698.75878
25	09/05/2014	Cavallo Morto	Corsica	41.4073888	9.17194444	70.30151
26	10/05/2014	Solenzara	Corsica	41.8628333	9.37988888	120.04943
28	10/05/2014	Bavella	Corsica	41.8106666	9.24669444	698.75878
30	11/05/2014	Asco	Corsica	42.4436111	9.01186111	733.60620
31	06/05/2014	Aritzo	Sardinia	39.9454166	9.19938888	1003.49414
32	12/05/2014	Pietralba	Corsica	42.5161111	9.17427777	282.27063

### **III.2.b DNA extraction, amplification, sequencing, and alignment**

We sequenced a total of 6 genes: a region of 658 bp of the *COI* for 347 individuals, as well as six nuclear genes (*wg*, *zen*, *ShxA*, *ShxB*, *ShxC* and *ShxD*) for a subset of 70 individuals spanning from north Africa to northern Europe. We further sequenced two outgroup sequences for the nuclear genes belonging to the closely related species *Pararge xiphioides* (Staudinger,

1871) (cf. Weingartner *et al.* 2006). For details on primers and cycling conditions see Appendix III – Note 1. The *COI* sequences were generated at the Biodiversity Institute of Ontario, Canada following standard protocols for DNA barcoding (deWaard *et al.* 2008), and DNA sequencing was performed on an ABI 3730xL capillary sequencer (Applied Biosystems).

### III.2.c Phylogenetic analyses

Evolutionary trees were constructed in MEGA7 (Kumar *et al.* 2016). The evolutionary history was inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model. The initial tree for the heuristic search was obtained automatically by applying the Maximum Parsimony method. The analysis involved 100 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 270 positions in the final dataset for *COI*. Sequences of *P. xiphioides*, *P. xiphia* and *L. megera* were recovered from NCBI (Weingartner *et al.*, 2006) and used as outgroup for *COI*. As outgroup for the *wg*, *zen* and *Shx* genes we used our recovered sequences from *P.xiphioides*, as well as the published *wg* sequences from Weingartner *et al.*, (2006).

### III.2.d Haplotype Networks and Genetic Landscapes

Patterns of genetic variation were analysed by inferring maximum parsimony haplotype networks using the program TCS 1.21, with a 95% connection limit (Clement *et al.*, 2000). Representations of genetic diversity were created for all of the investigated genes by calculating matrices of p-distances for each of them. Then, the dissimilarity matrices were projected in two dimensions by principal coordinate analysis (PCoA) using the „cmdscale“ R function. The bidimensional configurations for specimens were projected in RGB colour space using the same package (Dapporto *et al.*, 2014b). Specimens belonging to the same grid square of 2° for latitude and longitude were grouped, and their individual RGB colours were plotted on a Map.

Genetic variability among lineages over space has been computed for *COI* data as follows. The study area has been divided in squares of 0.2x0.2 degrees of latitude and longitude

and the specimens have been divided according to the phylogenetic tree as belonging to the Maghreb or to European lineage. Then we carried out two analyses for the two lineages separately. When more than two specimens were sequenced for each 0.2x0.2 squares, only two specimens were randomly selected for that square in order to reduce crowding of data. Then, for each lineage, the specimens examined for *COI* in a circular area of 200km of ray around each 0.2x0.2 square have been recorded. If a square had at least four specimens inside the 200km ray circle it was selected for the analysis. In cases with more than four specimens, the four nearest to the centre were selected and the others discarded. Then a p-distance matrix in *COI* sequences was calculated among the four selected specimens. Moreover, another matrix was calculated for each pair of specimens by summing their distances from the centre of the 0.2x0.2 square. Finally, the corresponding values of p-distances and sum of spatial distances have been divided resulting in a third dissimilarity matrix where genetic distances were weighted by their distances from the centre of the square. To return an overall value of genetic distance the values in this third matrix have been summed and the result multiplied for the mean value of summed geographic distances. This returned a single value for each spatial cell measuring the genetic differentiation of the four closest specimens weighted for their distance from the centre. These values were then imported into QGIS 2.0.1. (QGIS Development Team 2009), and interpolated using the inverse distance weighting algorithm to generate a visual representation of the spatial distribution of genetic variation.

### **III.2.e Pre-zygotic reproductive barriers: Courtship behaviour in Sardinian and Corsican *Pararge aegeria***

The crosses were performed with the offspring of the wild-caught females. The offspring resulting from the crosses (both hybrids and pure-bred Sardinian and Corsican individuals; F1), were crossed amongst each other (see below; i.e. backcrossed) to generate a F2 (see also Longdon *et al.* 2017). For those backcrosses no behavioural data was obtained. To perform the crosses, newly eclosed virgin females were placed in cages along with an artificial flower containing 10% honey solution (Goulson & Cory 1993; Gibbs *et al.* 2012). Newly

eclosed virgin males were then introduced and the total courtship duration (seconds) was timed using stopwatches. The cue used to determine the initiation of courtship was the bowed wing flick used by the males during the courtship ritual (Davies 1978). If the male was unsuccessful at initiating mating after numerous bouts of courtship between 8am and 6pm, then the male was removed and replaced with a new virgin male the following morning (8am). After mating had finished an egg laying plant was placed in the cage and the male was removed. Four types of crosses were set up: Corsican male/Corsican female (CC), Corsican male/Sardinian female (CS), Sardinian male/Sardinian female (SS) and Sardinian male/Corsican female (SC). In total 74 crosses generated data to be used in the analyses (CC=37, CS=17, SC=9, SS=11, Table 2).

**Table III - 2. Crosses used in the mating experiments.**

The numbers for each of the backcrosses, for which a successful mating was observed, with reference to the parents of the males and females used in the backcrosses (CC – animal had a mother and father from Corsica; SS – both parents from Sardinia; SC – father from Sardinia, mother from Corsica; CS – father from Corsica and mother from Sardinia).

<b>Cross type</b>		
Male	Female	Number of crosses
CC	CC	17
CC	CS	1
CC	SC	1
CS	CC	6
CS	SS	5
SC	CC	6
SC	SS	3
SS	CS	4
SS	SC	3
SS	SS	8

### **III.2.f Reproductive barriers**

After mating the female was left to oviposit for 6 days and all eggs laid in that period were collected. Female age throughout the experiments was recorded as it affects willingness to mate, and reproductive output (Bergman *et al.* 2010; Gibbs *et al.* 2010a; Gibbs *et al.* 2010b). Six days were allowed for oviposition as it represents the period of peak egg laying, usually followed by a rapid increase in mortality of both eggs and females (Gibbs *et al.* 2010b). After the 6 days females were removed and used for measurements of wing morphological traits. The first 8 larvae to hatch of a particular cross were reared through on a mix of *P. trivialis*, *D. glomerata*, *Brachypodium sylvaticum* and *Festuca rubra*. The hatching success of the remaining eggs was noted and the remaining larvae sacrificed. Larvae placed on food plants were followed through to eclosion and the proportion of individuals surviving to adulthood and the sex ratio of the adults was recorded. Pupae were sexed and kept individually, to ensure that virgin adults were available to set-up mating pairs in backcrosses.

After the individuals used in the crosses had been sacrificed their forewings were removed and the dorsal side of the forewing was placed in between glass slides and photographed using a Leica MZ6 dissection microscope with integrated camera (Leica IC80 HD camera with Las EZ software suite) under controlled light conditions. Wing area (mm<sup>2</sup>) of both forewings was measured through the use of imageJ software (Abramoff *et al.* 2004; Breuker *et al.* 2007; Breuker *et al.* 2010), and the average forewing area was used as a proximate measure of individuals' size (cf. Merckx & Van Dyck 2006), and included as a covariate in the models.

### **III.2.g Backcrosses**

In the backcrosses, females and males were kept in cages in conditions similar to those for the original crosses described above. For the backcrosses only hatching success of a sample of a minimum of 10 and a maximum of 20 eggs was assessed, as this was considered a representative sample size to determine hatching success. Only those crosses for which a successful mating was observed were included. A hybrid male, or female, was backcrossed to

either a purebred Sardinian or Corsican (Table III - 2). After the male had been removed females were provided with an egg laying plant and allowed to oviposit for 6 days (see also original crosses).

### **III.2.h *Wolbachia***

The wild-collected females whose offspring were used as parents in the crosses (with the exception of females 3 and 14; Table III - 1 and Appendix III. Table 1) were screened for the presence of *Wolbachia*, as this has been shown to sometimes affect reproductive output and fertility in insects, and the presence of this endosymbiont has been reported in *P. aegeria* ovaries (reviewed in Carter *et al.* 2013). This aimed to establish whether *Wolbachia* infections were present in the populations on Corsica and Sardinia used in this study. In order to screen for *Wolbachia*, we PCR amplified *Wolbachia* specific sequences (*Wolbachia* surface protein – *wsp*) using previously described primers (Dobson *et al.* 1999). The PCR products were run on a gel and screened for the presence of amplification. Individuals used in the crosses presented in this study were not tested for *Wolbachia* prior to mating as that was not feasible given the design of the experiments, nor postmating. The fact whether or not *Wolbachia* infection was detected in the mothers of the animals used to establish the crosses was used as a fixed factor in the models described below.

### **III.2.i Statistical analyses Mating Experiments**

Linear mixed effect models (fitted by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom) were constructed to investigate variability amongst the crosses in reproductive output (both numbers of eggs laid and egg hatching success), larval survival and courtship duration. The latter is the net result of the choosiness of the female, and the willingness and effectiveness of the male. Minimum models were constructed using AIC as a guideline, and these are the models presented in this study. This means that non-significant fixed covariates and interactions were removed. Once model selection had been completed, significance of the remaining fixed effects was provided through use of the lmerTest package

(Kuznetsova *et al.* 2016). Fixed effects tested for inclusion were age of both male and female at the time of mating, their size (measured as wing size), *Wolbachia* infection (see aforementioned) and type of cross. The R-packages used Satterthwaite approximations to estimate the denominator degrees of freedom and can provide *p*-values for type III ANOVAs to display the significance of the fixed effects; these are the values that are presented in the text. All residuals for included effects were tested for normality and log and square root transformations were used where appropriate (e.g. courtship duration). Both male and female maternal origin were kept as random factors in all the models, and as the models tested the significance of differences between the various cross types, cross type was always included as a fixed effect. Analyses of the mating data, but also of the Haplotype Networks and Genetic Landscapes (details in section III.2.d) were performed using R (3.4.0) (packages: lme4, rcmdr, lmerTest) (R Development Core Team 2016). Chi-square tests were used to test for cross type and fertility associations; while for the backcrosses the Fisher's Exact Test for Count Data, as some counts were very low (see Table III-2).

### **III.3. Results**

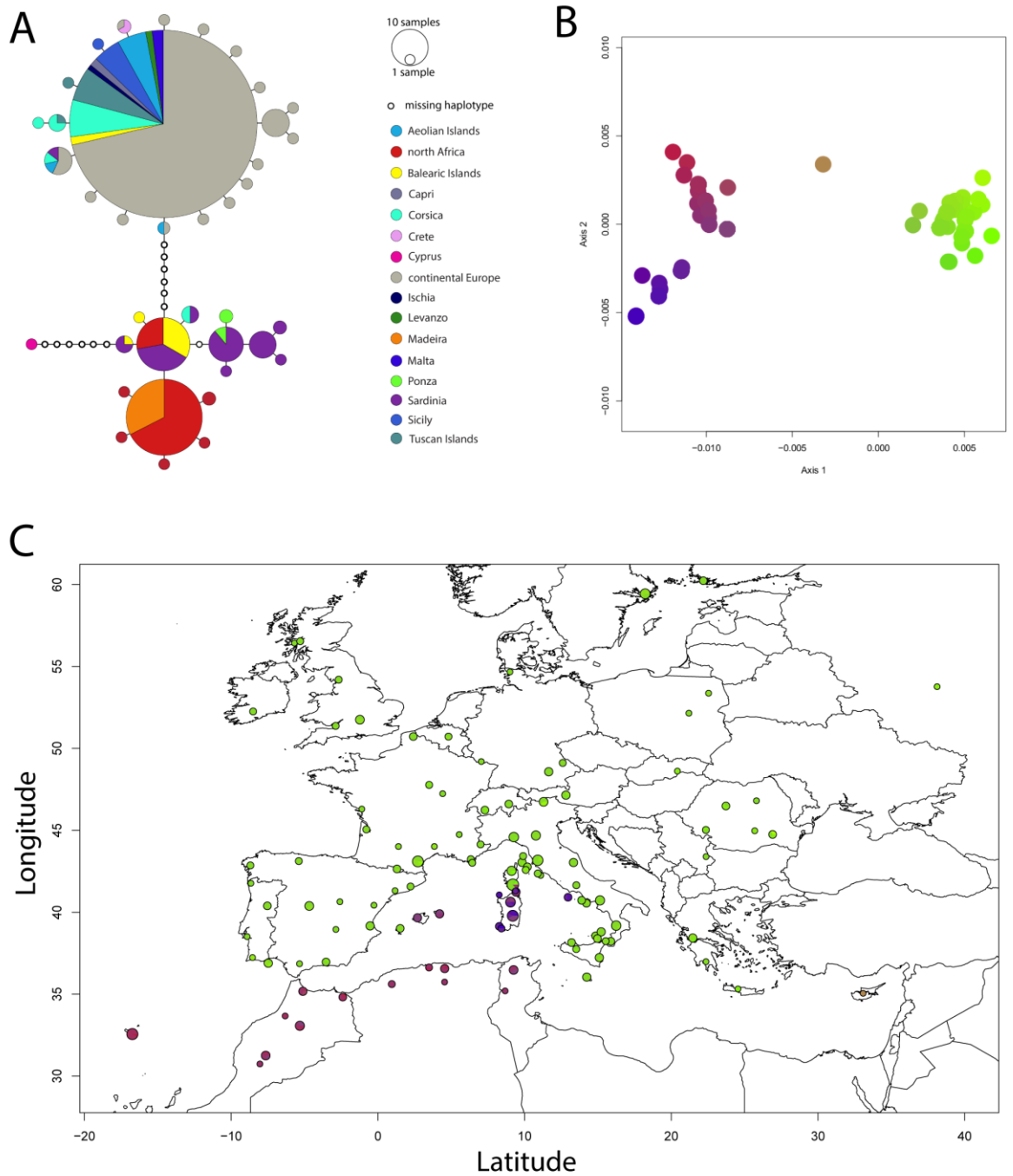
#### **III.3.a *COI* variation reveals the presence of two distinct lineages**

We obtained 346 sequences with 27 haplotypes characterized by 28 variable nucleotide sites for the *COI* gene. Haplotype networks based on *COI* sequences show a discrete boundary between North African and European populations, forming two distinct lineages separated by a minimum of 6 mutations (Figure III - 1). North African haplotypes show significant population structure, with several highly frequent haplotypes being present throughout the areas analysed. In contrast, populations in continental Europe are characterised by one main haplotype, connected to several low frequency ones by a maximum of two mutations (Figure III - 1). Interestingly, the islands of Sardinia, Mallorca, Menorca and Ponza were all populated exclusively by North African haplotypes, even though they are in closer proximity to continental Europe (Figure III – 1, **B**). Furthermore, we found evidence of only one individual

carrying the Sardinian haplotype in Corsica (Bonifacio), suggesting a very limited gene flow between the two islands.

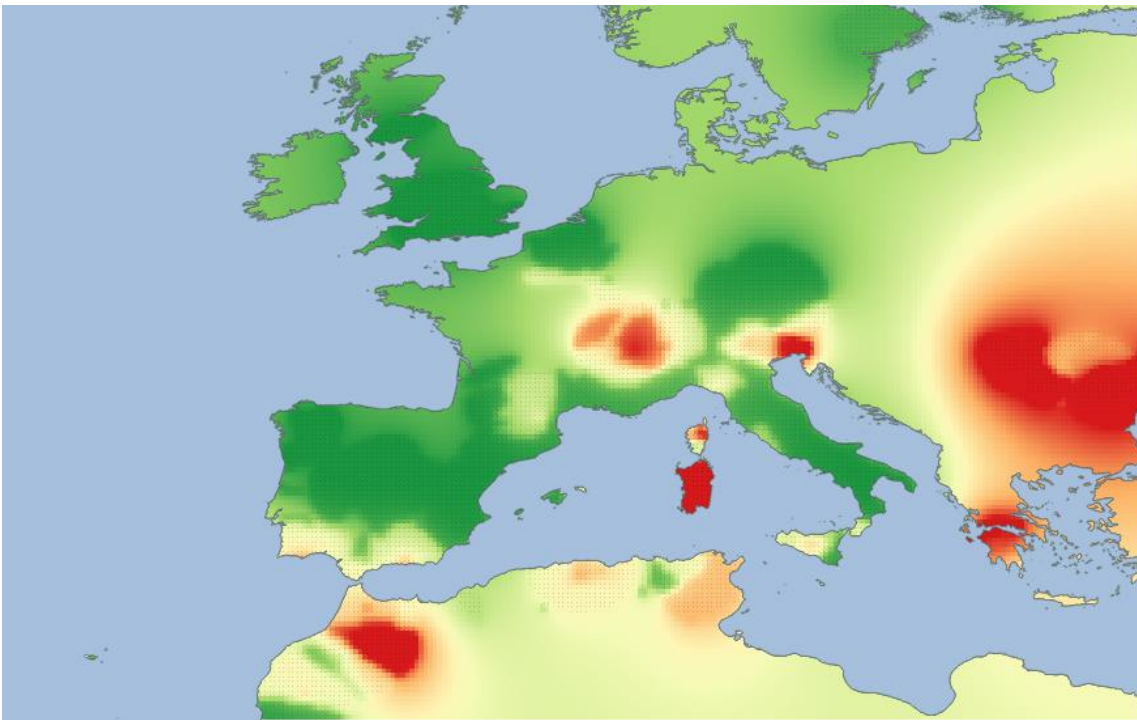
When splitting the populations based on the *COI* lineages, we observed a significant negative Tajima's D for the European lineage (Tajima's D = -2.10,  $p < 0.05$ ), but not for the North African one (Tajima's D = -0.91,  $p > 0.10$ ; Combined Tajima's D = -0.49,  $p > 0.10$ ). Overall genetic diversity was also higher for the North African lineage as compared to the European populations (average nucleotide diversity,  $\pi$  was 0.0025 and 0.0013 respectively). This is also evident when the genetic differences among the nearest 4 specimens to each  $0.2 \times 0.2$  square of latitude and longitude is plotted on a map (Figure III - 2). Geographical locations corresponding to the North African lineage are shown to harbour more genetic heterogeneity. Interestingly, the populations in Romania and Alps are also more variable, which suggests increased genetic diversity for the European clade in Central and Eastern Europe.





**Figure III - 1. Distribution of *COI* haplotypes across Europe.**

**A**- Haplotype network constructed under a TCS framework. Each colour indicates a different haplotype where the size of the circle corresponds to the frequency of a haplotype. The number of nucleotide changes at each node is shown as white circles (putative ancestral haplotypes). **B** – Pcoa of genetic distance matrix based on *COI* sequences. **C** – Projection of the Pcoa onto geographic space showing *COI* clustering.

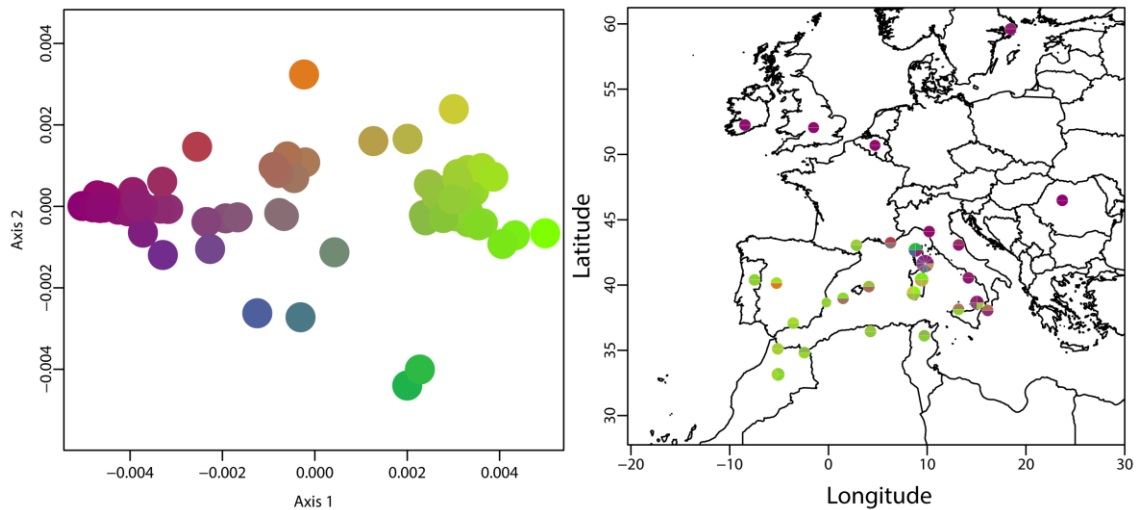


**Figure III - 2. Genetic Heterogeneity of *COI* in *Pararge aegeria*.**

Genetic richness based on 0.25x0.25 degree squares for which there are at least 4 specimens sequenced in the 100km radius. Genetic richness was calculated separately for the two lineages (in haplotype divergence) for each of these squares, represented here as a heat map.

### **III.3.b Nuclear genes versus *COI* lineages**

When using a concatenated dataset, sequence variation in all of the nuclear genes combined (i.e concatenated) revealed to be in slight discordance with the mtDNA pattern since the pattern of genetic clustering showed a south-western genotype mainly distributed across north Africa, Iberia, France, Sardinia and Sicily and a north-eastern genotype in the Italian Peninsula, north Europe and Eastern Europe (Figure III - 3). As a main difference between the spatial pattern of mitochondrial and nuclear genes, the Iberian Peninsula and Sicily are inhabited solely by *COI* haplotypes belonging to the European lineage, while nuclear sequences also belong to the southern-western lineage.

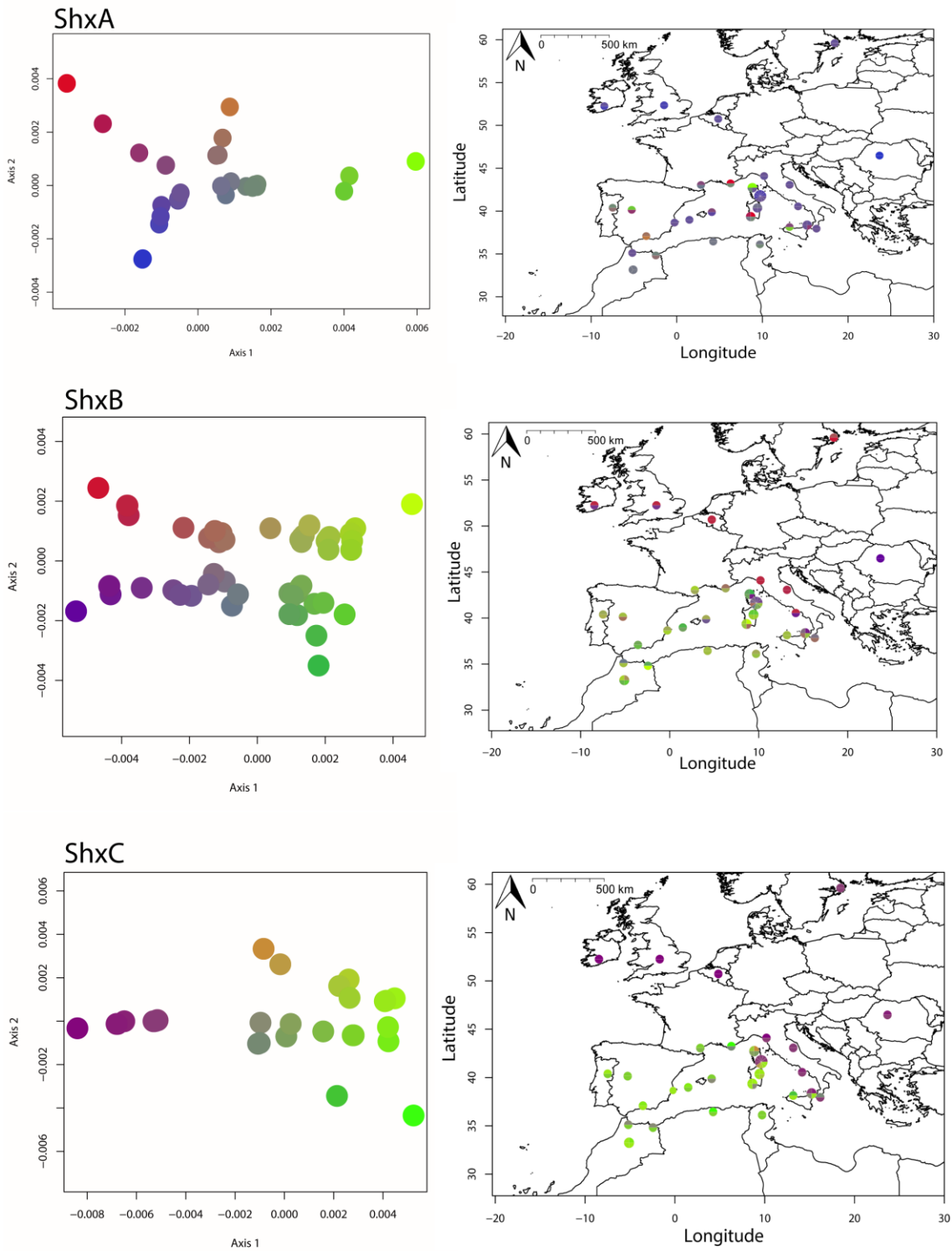


**Figure III - 3. Geographic clustering of nuclear genes in *Pararge aegeria*.**

(A) P-distance PcoA projected in RGB colour space based on the concatenated nuclear dataset. Most individuals cluster as two distinct groups, with some intermediate genotypes. (B) Geographical location of populations with RGB colours as displayed in A.

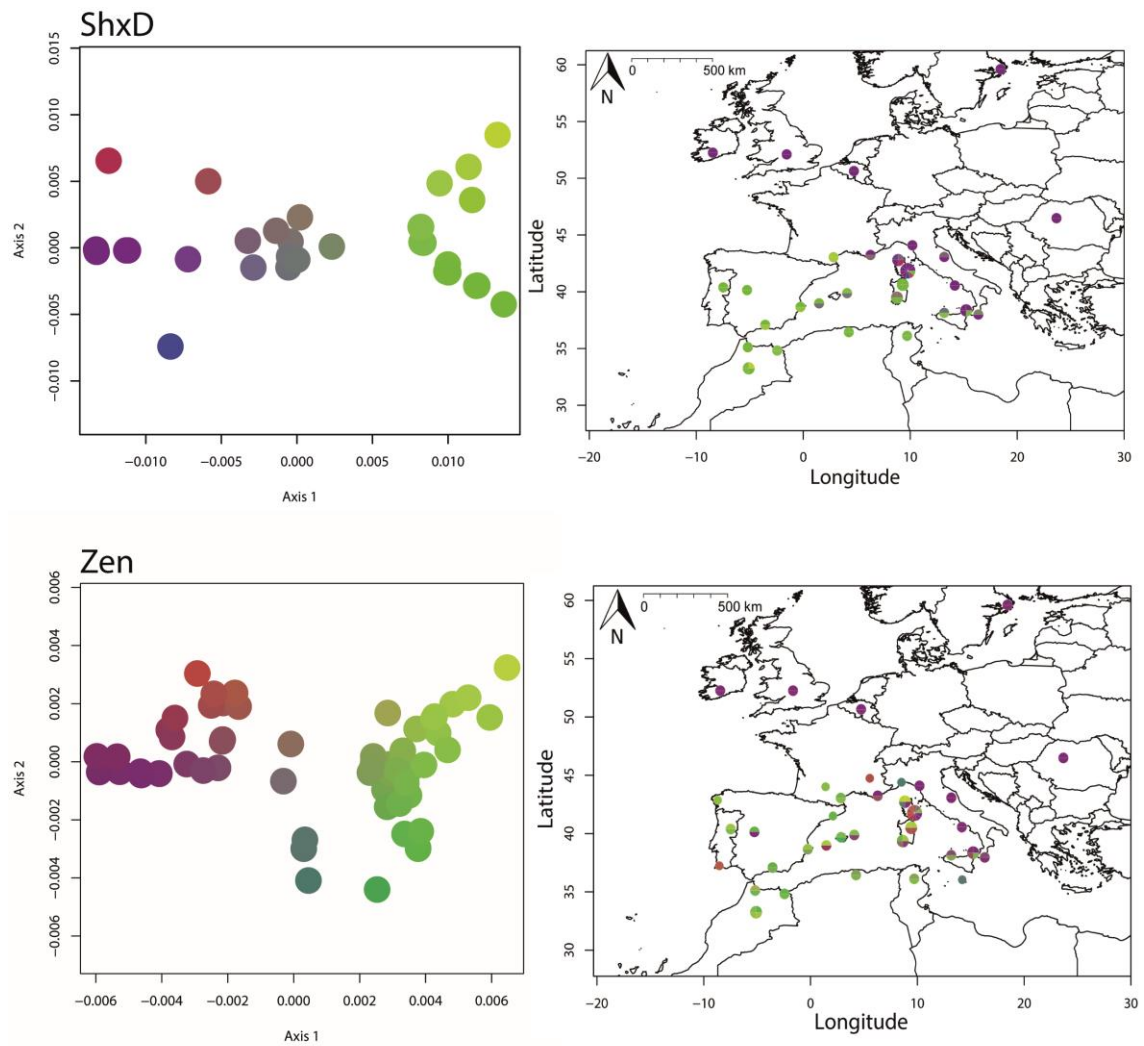
### III.3.c Patterns of polymorphisms associated with the *Shx* genes

As discussed in Chapter II, several polymorphisms were found segregating within *P. aegeria* populations within the different *Hox3* genes. Differential patterns of clustered polymorphisms as compared to *COI* and between the *Shx* genes could be indicative of differential selection pressures operating on individual genes. Overall, *ShxA* showed the most homogenous distribution, with little geographic structure (Figure III - IV). This is likely due to low levels of divergence within *ShxA* (Chapter II – Table II -4). *ShxB* showed more geographic structuring, similar to the concatenated dataset, where similar sequences cluster in a south-west/north-east manner. Similar patterns are observed for *ShxC*, *ShxD* and *zen*. Interestingly, several high frequency homozygous polymorphisms were found segregating within the homeodomain of *ShxD* (See Chapter II). Of note is an Ile to Met substitution occurring at position 55 of the homeodomain in 36 of the sampled individuals. This is a known position within the third alpha helix known to make DNA contact (Hanes and Brent, 1989), and segregates in a similar geographic pattern to other polymorphisms in the *Hox3* genes.



**Figure IV – 4. Differential patterns of polymorphisms associated with the *Hox3* locus in *Pararge aegeria*.**

P-distance PcoAs projected in RGB colour space based on the individual *Hox3* genes shown on the left. On the right, the individual points are projected onto a map covering the geographical range sampled. Continued on the next page.



**Figure IV – 4 Continued.**

### **III.3.d Pre-zygotic reproductive barriers: Courtship behaviour**

Females in pure-bred Corsican crosses were significantly slower in mating than any of the other crosses (full minimum mixed model AIC=142.1, BIC=156.6, d.f. resid = 52). Not only did they take longer to mate compared to Sardinian females in pure-bred Sardinian crosses (CC versus SS;  $t=-2.80$ ,  $df=59$ ,  $p=0.0068$ ), but Sardinian females also mated more readily with Corsican males, than Corsican females did (CC versus CS  $t=-3.61$ ,  $df=59$ ,  $p<<0.001$ ). Finally, Sardinian males also mated more readily with Corsican females, than Corsican males did (CC versus SC  $t=-2.18$ ,  $df=59$ ,  $p=0.033$ ). In terms of mating opportunities, Corsican pure-breds

could be outcompeted by Sardinians following dispersal. Female age, size or temperature did not improve the model.

### III.3.e Reproductive barriers

*Female fecundity:* Reproductive output (i.e. number of eggs laid) was significantly affected by female age and size, as well as cross type (AIC=605.8, BIC=626.3, d.f. resid = 63). Females that were older at the time of mating laid more eggs in the 6 days following mating than those that mated young, having presumably stored mature eggs for fertilisation ( $t=3.25$ ,  $df=71.90$ ,  $p=0.0018$ ). Larger females laid significantly more eggs ( $t=2.88$ ,  $df=67.48$ ,  $p=0.0053$ ). Sardinian females (i.e. the SS and CS crosses) laid significantly fewer eggs than Corsican females (i.e. the CC and SC crosses), regardless of the origin of the male to whom they mated (SS versus CC  $t=-3.31$ ,  $df=20.53$ ,  $p=0.0034$ ; CS versus CC  $t=-3.87$ ,  $df=15.23$ ,  $p=0.0015$ ). There was no significant difference between CC and SC ( $t=-1.75$ ,  $df=71.82$ ,  $p=0.085$ ).

*Offspring fitness and the effect of temperature on egg hatching success:* All four types of crosses were similar in terms of infertile (i.e. egg hatching success =0%, or no eggs laid, despite having been observed to mate successfully) versus fertile (i.e. egg hatching success > 0%) crosses (chi-square 1.58,  $df=3$ , and  $p=0.66$ ). Corsican and Sardinian *P. aegeria* can thus interbreed freely, without significant fertility issues (i.e. ratio's fertile and infertile crosses were the same). Indeed, there were no significant differences in egg hatching between the different cross types, with egg hatching success only affected by temperature, but not female age at mating or female size (AIC=506.6, BIC=523.9, d.f. resid=57). Within the temperature range used (range: 22.1 – 25.4°C), more eggs hatched successfully at higher temperatures ( $t=2.43$ , d.f. =60.82,  $p=0.018$ ).

There were no significant differences (i.e.  $P \gg 0.05$ ) in survival of the offspring (i.e. from larval hatching to eclosion as an adult) between the crosses (full model with only cross type AIC=32.7, BIC=48.0, d.f. resid=58).

*Wolbachia* infection status: The majority of the field-collected females were found to be infected with *Wolbachia*, with the exception of five females, three from Aritzo (Sardinia), one from Desulo (Sardinia), and one from Bonifacio (Corsica). However, Aritzo is not a location free from *Wolbachia*, as other females collected there were found to be infected (Sup. Table 2). We cannot rule out *Wolbachia* presence in populations from Desulo and Bonifacio as only a single specimen was collected in each of these localities. The *Wolbachia* infection status of mothers was not a factor that significantly improved the statistical models reported earlier, and therefore was not included in the reported final models. Finally, for each of the four cross types Chi-squared tests were used to evaluate the presence of sex ratio distortion in the surviving offspring. No significant sex ratio distortion was found in any of cross types: CC (chi-square=0.12, df=1, p=0.73), CS (chi-square = 0.017, df=1, p=0.90), SC (chi-square = 0.059, df=1, p=0.81) or SS (chi-square = 0.76, df=1, p=0.78). The lack of sex ratio distortion and the absence of any obvious fertility problems suggest that the presence/absence of *Wolbachia* is unlikely to explain variation in reproductive output across cross types.

### **III.3.f Post-zygotic reproductive barriers**

*Sterility of F<sub>1</sub> hybrids*: F1 hybrids were backcrossed to either pure-bred Sardinians or Corsicans (Table 2). There were no differences between the 10 types of crosses in terms of fertility (Correcting for low counts; Fisher's Exact Test for Count Data, p=0.13).

## **III.4. Discussion**

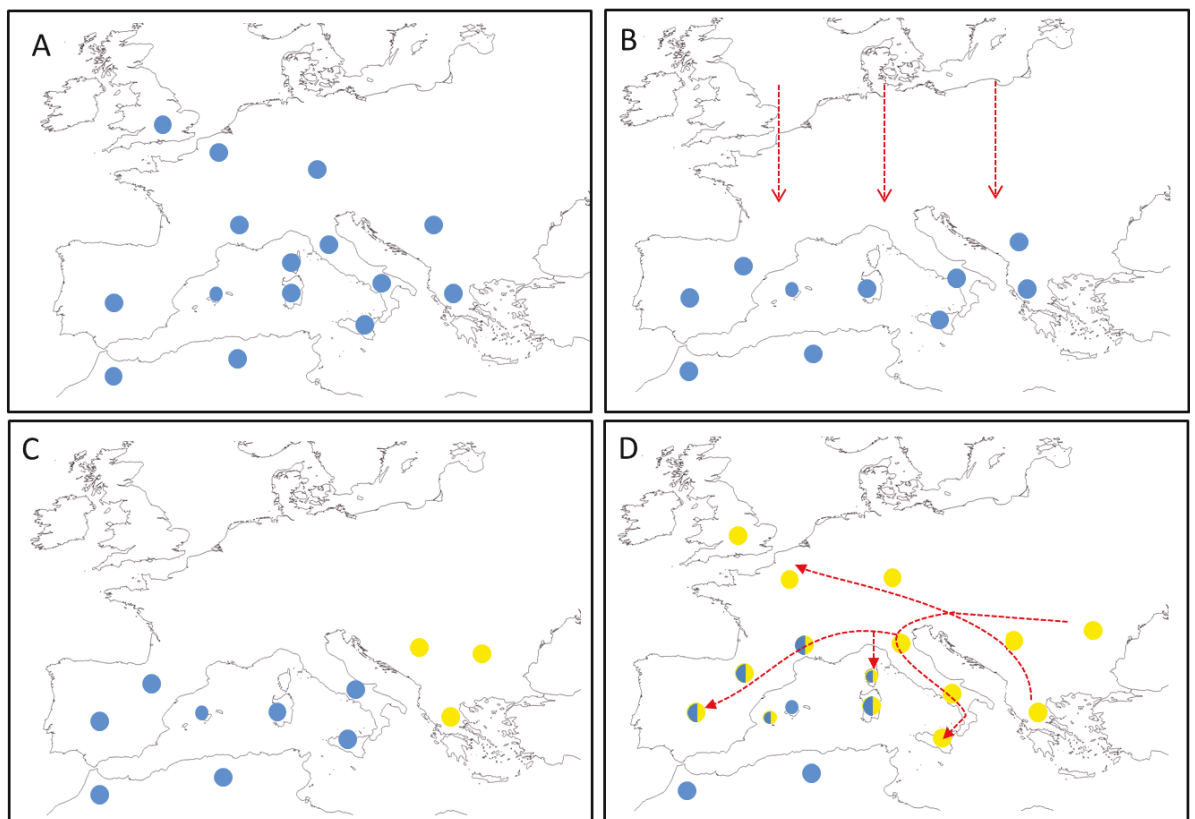
Corsica and Sardinia are characterised by the occurrence of a variety of endemic populations (Grill *et al.* 2002; Dapporto & Dennis 2009; Dapporto 2010). This is likely to be the result of the long-term isolation of these islands since the early or late Miocene (Alvarez 1972; Mauffret *et al.* 1999; Meulenkamp & Sissingh 2003; Ketmaier *et al.* 2006). Mutually exclusive pairs of cryptic butterfly species such as *Aricia agestis* and *A. cramera* or *Polyommatus icarus* and *P. celina* have been shown to occur on Corsica and Sardinia (Dincă *et al.* 2011, Vodă *et al.*

2015a, 2015b), and lack of gene flow has previously been inferred for *P. aegeria* (Dapporto 2010). Such divergence between Corsica and Sardinia populations is in many ways unexpected as they are in close proximity, separated by only a small sea strait (11 km wide), and even share a biogeographic history to the onset of the Holocene (Dapporto 2010, van der Geer et al. 2010). In Sweden, *P. aegeria* revealed little to no gene flow between the populations of the island Öland and the near-by mainland (5km) (Tison *et al.* 2014). The data presented in this study confirm that even short sea straits can provide a strong barrier to the dispersal of *P. aegeria*. However, discordant patterns were observed between the nuclear genes (*wg* and *Hox3* genes) and *COI*, where the Iberian Peninsula is inhabited solely by *COI* haplotypes belonging to the European lineage, but the nuclear genes clustered geographically between North Africa and Iberia, as well as Sicily. Nuclear markers also revealed extensive sharing of haplotypes between the islands of Corsica and Sardinia. This discordance could indicate different rates of genetic flow between nuclear and mitochondrial genes, either because of a difference in rates by which the nuclear genes (both *wg* and *Hox3* genes) evolve (as a result of differences in the rates of incurring mutations, possibly in combination with differences in fitness associated with *wg/Hox3* and *COI*), and/or male-female biased dispersal (Toews & Brelsford 2012).

The presence of the north African *COI* lineage on the Mediterranean islands is intriguing, as they are in closer proximity to the European mainland in terms of length of sea straits, and one would expect them to be more easily colonised from either the Italian Peninsula (in the case of Ponza and Sardinia) or the Iberian Peninsula (in the case of Mallorca and Menorca). The higher genetic heterogeneity observed in the North African lineage (Figure III - 2), suggests the presence of ancestral populations in North Africa and the associated islands. This is in stark contrast to the reduced genetic variation observed in the European clade in the circum-Mediterranean populations, suggestive of a recent colonisation and population expansion. This may also explain the overall observed results presented in Chapter II for the *Hox3* genes, where largely negative Tajima's D-values were reported across the range of *P. aegeria*, including Europe. If that is the case, then if *COI* is evolving faster than the nuclear genes anyway, it could create a discordance between the distribution of polymorphisms in the



nuclear genes and *COI*. The significant negative Tajima's D for European populations also supports this hypothesis, where low frequency variants segregate at high frequencies, which could be indicative of population expansion following a bottleneck or selective sweeps (Waters, 2011). Given the higher genetic variation found on Alps and in Romania (Figure III - 2) one could propose a putative centre of origin for the European populations further east, which colonised and outcompeted the ancestral populations in Western Europe (Figure III - 4). This could have occurred with a phalanx-like colonisation over the mainland (see also Figure III - 5), which was impeded by where the island lineages tend to be unexpectedly similar to the N. African populations (Dapporto et al. 2012). The populations in Sardinia, Mallorca, Menorca and Ponza might thus represent "relict" populations harbouring the ancestral *COI* haplotypes, which have not been replaced due to the physical barriers imposed by the sea straits.



**Figure III - 5. Possible recolonisation scenario for *Pararge aegeria*.**

The ancestral *COI* haplotype (blue circles) was present throughout the range of *P. aegeria* in Europe (A), without differentiation of the nDNA genes due to unrestricted dispersal between populations. During the last glacial period (B) the range retracted southwards (red arrows), and gene flow was restricted between the refugia due to the Alps and Pyrenees acting as barriers, which allowed for periods of differentiation (yellow circles in C). Following the warming of the climate, populations carrying the supposedly advantageous *COI* variant spread northwards and westwards (red arrows in D), where it could have admixed with the nuclear genome of warm adapted populations in the Iberian peninsula as well as the islands of Ibiza, Corsica and Sardinia resulting in the discordance between the markers (indicated by blue and yellow circles in D). This admixture was presumably hindered by sea straits, giving rise to the sharp boundary observed for the *COI*.

However, it must be noted that *COI* is maternally inherited and it can only trace the dynamics of females. Nuclear genes show a general correspondence into two main southern and northern groups but also areas of discrepancy where the northern *COI* is associated to southern *wg/Hox3* genes. The data presented in this chapter show that hybrid sterility and hybrid-purebred incompatibilities do not limit admixture between these islands, and there appear to be no obvious pre- or postzygotic barriers. Moreover, we observed that the two *COI* lineages are highly inter-fertile, but also that there are temperature-related differences across types in both female fecundity and offspring fitness during the egg stage hinting at possible effects of local adaptation to the temperature during oviposition and embryogenesis. All these data are indicative of a strong complexity and indicate that emergence and maintenance of a neat population structuring may be the result of strong selection on certain traits related to oogenesis and early development, of biased dispersal between sexes and/or genetic variants and genes in different parts of the *P. aegeria* range. Other Speckled wood populations across Europe show significant and distinct population structuring, evidenced by sequence analyses of the *P. aegeria* specific Rhabdovirus PAegRV (Longdon *et al.* 2017) and population genetic analyses (Tison *et al.* 2014). For the UK in particular, this is remarkable, given the relatively recent contraction and subsequent expansion of *P. aegeria* in the UK (Hill *et al.* 1999a; Longdon *et al.* 2017). Rather interestingly, PAegRV sequence variation indicates not only limited gene flow from Sardinia to south Corsica, but also that the Corsican PAegRV is basal to Corsican and Sardinian PaegRV, and thus that gene flow between the islands may not always have been so strongly limited. Strong differences between *P. aegeria* populations are not only evident on the basis of sequence variation, but also in terms of gene expression patterns and even on the basis of the presence of unique (miRNA) genes (Quah *et al.* 2015). This has been shown for egg production in Corsican (specifically Zonza) and Belgium populations. The results of Quah *et al.* (2015) also lead one to hypothesise that female reproductive strategies, and the genes involved therein, are very likely to be under selection in response to habitat variation (e.g. temperature and oviposition plants) with significant population differences, as observed in other *P. aegeria*

populations across Europe (Gibbs & Van Dyck 2009; Gibbs *et al.* 2010b; Gibbs & Van Dyck 2010).

A further factor that may be influencing the observed patterns of variation may be the presence of the detected *Wolbachia* harboured by the populations on both islands. While infection status did not considerably improve the mating models employed in this study, the primers used for detection of the symbiont did not distinguish between different strains. It is therefore possible that more subtle effects may be present as a result of the infection which may be influencing the observed patterns of genetic variation segregating between the two islands. For example, *Wolbachia* infection has been linked to viral resistance in *Drosophila* and several species of mosquitoes (Teixeira *et al.*, 2008; Hedges *et al.*, 2008). It is therefore possible that a particular strain of *Wolbachia* may be influencing the fitness of populations by conferring resistance to endemic viruses and/or pesticides. Further testing would be required to study the contribution of strains to the establishment of genetic differentiation in this model.

Overall, we observe a clear pattern in terms of *COI* haplotypes, where the Mediterranean provides a clear barrier between the two main lineages, suggesting limited dispersal of females. However, the sharing of nuclear polymorphisms across the Maghreb and Spain, as well as between the Mediterranean islands, suggests that males are able to migrate across sea straits and successfully reproduce between *COI* lineages. This pattern is reinforced by the mating data recovered in this study, where we observed that both lineages are highly interfertile, although Sardinian females mate more quickly with both local and Corsican males. This suggests that in terms of mating opportunities, Corsican pure-breds could be outcompeted by Sardinians following dispersal, but not vice-versa. It appears that the sea straits provide a barrier to female dispersal, while males appear to be able to cross and can disperse more readily, thus leading to the discordant patterns observed in the nuclear genes. These discordant patterns are reinforced by the geometric morphometric split observed for male genitalia shape between populations of *P. aegeria* (Dapporto *et al.* 2012), where the same south-west/north-east differentiation pattern is observed.

Another possibility is that the areas of discordance arose as a result of adaptive polymorphisms being retained in the areas of discordance, which in itself is not incompatible with the fact that the *Hox3* and *wg* may be evolving slower than *COI*. Chapter II did show them to be under largely negative selection. As has been hypothesised for *COI*, it is possible that more adaptive *Shx* variants could have been admixed into populations colonising from Eastern Europe, and retained due to potential advantageous functions. However, this is unlikely for various reasons. Firstly, the same differentiation pattern is also observed for the unrelated developmental gene *wg*, which has likely experienced different selective pressures to the *Shx* genes. Furthermore, the differentiation patterns observed for the *Shx* genes match several other patterns which have been observed for other unrelated markers in other species (Schmitt, 2007). It is more likely, therefore, that the differentiation signals come from neutral variation that accumulated during periods of isolation during the last glacial maxima, and that nuclear genes accumulate such variation at slower rates from *COI*. In that respect the distribution of *ShxA* is of interest, displaying a rather uniform distribution of its polymorphisms, and having low genetic diversity.

The data add support to the hypothesised recolonisation scenario outlined in Figure III-5, but suggest that the recolonisation over land depicted by red arrows in Figure III-5D may be largely driven by male dispersal. This is an interesting finding because female Speckled Woods are often considered to be the most dispersive sex in a more metapopulation setting (Hughes *et al.*, 2003). However, evidence also exists suggesting that dispersal is more costly to females, often lowering reproductive output (Hughes *et al.*, 2003; Bergerot *et al.*, 2012). It is therefore possible that, while females do indeed disperse, the cost associated with moving across sea straits is heavily penalised in females, even more so than in a metapopulation setting, and so maternal *COI* does not make it across sea straits. On the other hand, even though male dispersal is less common in *P aegeria* males, reproductive output may not be affected.

## Chapter IV

# Development of CRISPR/Cas9 technology in *Pararge aegeria*

# Abstract

With the exception of a few moth and butterfly species, gene editing tools are lagging behind other organisms in the Lepidoptera. In order to study gene function across the order, it is necessary to establish tools that enable gene manipulation. CRISPR/Cas9 is a promising technique that has been very effective in a wide range of species, including several butterfly species. However, no attempt has yet been made at establishing this technique for my model species *Pararge aegeria*. With this in mind, I aimed to initially establish the technique through the targeting of the wing patterning genes *yellow* and *WntA*. Injection of sgRNAs targeted at these loci confirmed the effectiveness of this technique in *P. aegeria*, and resulted in phenotypes consistent with their hypothesised roles in wing development. I then used CRISPR/Cas9 to target two *Shx* paralogs. I was able to show that injections resulted in large deletions affecting the loci, but could not recover any mutant phenotypes. I discuss further applications of the technique, especially in relation to the study of early developmental genes.

## IV.1. Introduction

The developmental evolution of butterfly wing patterns is a major model for the evolution of morphological novelties and the co-option of particular developmental genes therein (Mazo-vargas *et al.*, 2017). Furthermore, several recent studies have highlighted the divergent patterns of early embryogenesis, and maternal gene regulation in moths (e.g. *Bombyx mori*) and butterflies (e.g. *Pararge aegeria*) (Carter *et al.*, 2013; Carter *et al.*, 2015; Ferguson *et al.*, 2014; Nakao, 2010, 2012, 2016; Nakao *et al.*, 2008). Although these studies have identified key genes involved, and elucidated their expression patterns and sequence variability (Ferguson *et al.*, 2014) (see also Chapter II), we thus far lack the tools to adequately study their function. While some progress has been made for *B. mori*, Lepidoptera are somewhat unique as a major model system lacking such functional tools. In the first instance, Lepidoptera appear to have a very poor response to RNAi (Terenius *et al.*, 2011; Kollipoulou and Swevers, 2014) which some have argued is a result of high levels of expression of dsRNases present endogenously within the order (Shukla *et al.*, 2016), as well as poor intracellular transport of the injected RNA (Joga *et al.*, 2016). As a result, the biological function of most genes within the order remains unknown. Although some tools have been developed specifically for the species *Bicyclus anynana* (Chen *et al.*, 2011; Marcus *et al.*, 2004; Monteiro *et al.*, 2013; Ramos and Monteiro, 2007; Ramos *et al.*, 2006; Tong *et al.*, 2014), there remains a strong need to establish genetic engineering methods that may be applicable to various lepidopteran species, in order to start elucidating gene function across the order.

Several techniques are available that enable the study of gene function in insects through knockout/knockdown. Zinc finger nucleases (ZFNs) and activator-like effector nucleases (TALENs) for example, have both been shown to work for *B. mori* (Takasu *et al.*, 2010; Takasu *et al.*, 2013; Takasu *et al.*, 2016). However, both ZFNs and TALENs require the engineering of nucleases, which is time consuming, costly and technically challenging (Gaj *et al.*, 2013). Furthermore, while a retrotransposon system using PiggyBac vectors has been shown to work in *B. anynana* (Marcus *et al.*, 2004) and *B. mori* (Tamura *et al.*, 2000), it

presents a random genomic insertion problem as it does not allow for targeted genomic integration.

The clustered regularly interspaced short palindromic repeat/CRISPR-associated (CRISPR/Cas9) system has recently been developed to study gene function. The CRISPR/Cas9 system is an RNA-guided nuclease tool for the targeted introduction of double stranded DNA cleavage (Harrison *et al.*, 2014; Jinek *et al.*, 2014), which has been shown to work across a very wide range of organisms including nematodes (Friedland *et al.*, 2013), zebrafish (Hwang *et al.*, 2013), fruit flies (Gratz *et al.*, 2013; Gratz *et al.*, 2015) and more recently several species of moths (Bi *et al.*, 2016; Koutroumpa *et al.*, 2016; Wang *et al.*, 2013) and butterflies (Fujiwara and Nishikawa, 2016; Markert *et al.*, 2016; Perry *et al.*, 2016; Zhang and Reed, 2016; Zhang *et al.*, 2017; Mazo-vargas *et al.*, 2017). To knockout expression of the target genes, a synthetic guide RNA (sgRNA) is introduced into the desired organism, in conjunction with Cas9 protein or plasmid/capped mRNA encoding the protein. A 20nt region in the sgRNA complementary to a portion of the CDS of the target gene binds the DNA and directs the recruitment of Cas9, an endonuclease which induces double strand breaks at the desired site (Hsu *et al.*, 2014). These double-strand breaks result in indels at the targeted loci as a result of ineffective repair by non-homologous end joining, resulting in non-functional proteins in those cells carrying the mutations.

The power of the CRISPR/Cas9 system as compared to other gene editing methods lies in its simplicity: a single sgRNA is sufficient to guide the binding of Cas9, and the N<sub>20</sub>NGG (PAM) motif required for this binding is readily available within most target sequences. However, this technique isn't without its limitations. In particular, and as encountered during its implementation discussed in this chapter, it poses significant challenges for its application to genes that regulate key developmental processes during early embryogenesis (Harrison *et al.*, 2014). Introducing mutations early in development for such key genes is likely to be lethal to the animal before it reaches later developmental stages. Furthermore, interpreting mosaic phenotypes resulting from these injections is often challenging, as they result in partial



phenotypes in the tissue of interest. These limitations can be overcome by the development of transgenic animals containing conditional alleles, which allows for spatial and temporal control over gene knockout (Harrison *et al.*, 2014). The CRISPR/Cas9 system can also be exploited for this purpose, through the introduction of expression cassettes which allow for temporal control of their expression through, for example, a heat shock activated element (Waijers *et al.*, 2013). The system has indeed been used for the generation of F<sub>0</sub> knock-ins in mice (Platt *et al.*, 2014), zebrafish (Auer *et al.*, 2014) and fruit flies (Gratz *et al.*, 2013), amongst other organisms.

The technique of CRISPR/Cas9 has thus several limitations, and poses several (technical) challenges, but at present it is the most promising technique available to facilitate the study of the functionality of developmental genes in butterflies. The developmental genes and their interactions underpinning wing patterning are at present the best characterised aspect of development in butterflies (For reviews see for e.g. Jiggins *et al.*, 2017; Monteiro, 2015; Skelhorn *et al.*, 2016). Optimising the technique for my model species, *P. aegeria*, thus focused in the first instance on key wing patterning genes. In the second instance, a major aim is to use CRISPR/Cas9 to study the functionality of genes involved in the maternal regulation of early butterfly embryogenesis and elucidate the nature of the divergent patterns of development observed. The key genes to be targeted here were the *Shx* genes, which not only evolve rapidly (see Chapter II), but also appear to be pleiotropic: they are involved in specifying the extraembryonic serosa (Figure I – 4) (Ferguson *et al.*, 2014), and in later development also appear to be expressed in embryonic tissues. Developing functional tools to study these genes will allow us to determine their patterns of sub- and neofunctionalisation and evolution (see Introductory Chapter I).

For the wing patterning genes, the first set of genes to be investigated were those involved in the biosynthesis of pigments in insects (Wittkopp and Beldade, 2009), and whose *cis*-regulatory modules have been implicated in the evolution of pigmentation in *Drosophila* (Arnoult *et al.*, 2013; Gompel *et al.*, 2005; Prud'homme *et al.*, 2006; Rebeiz and Williams, 2017). The *yellow* genes comprise a family of rather divergent genes, involved in a number of

developmental pathways (i.e. are also pleiotropic), including melanisation (e.g. *yellow-y*; see Ferguson *et al.*, 2011). The ortholog of *yellow*, was recovered from the *P. aegeria* genome for CRISPR/Cas9 treatment. Furthermore, an ortholog of this gene has also been targeted in other butterfly species using the CRISPR/Cas9 system, and has been shown to result in mosaic individuals carrying mutations for the genes in the F<sub>0</sub>, resulting in pigmentation phenotypes (Perry *et al.*, 2016; Zhang *et al.*, 2017). In a recent paper, Zhang *et al.* (2017) were able to introduce mutations for several pigmentation genes in four different butterfly species, outlining the versatility of the CRISPR/Cas9 system in its application to diverse species.

In addition to the genes involved in pigmentation *per se*, the Wnt ligand *WntA* was also targeted, which has been shown to be involved in regulating wing patterning across a wide range of butterfly species (Belleghem *et al.*, 2017; Gallant *et al.*, 2014; Jiggins *et al.*, 2017; Kronforst and Papa, 2015; Martin and Reed, 2014). Although *WntA* is in all likelihood co-opted into butterfly wing patterning, ironically little is actually known about the (pleiotropic) functionality of *WntA* outside Lepidoptera. It is even absent from the *Drosophila* genome (Bolognesi *et al.*, 2008). In *Tribolium* at least, it has a significant patterning role in various aspects of early embryogenesis, which may be associated with short-germ development and growth (Bolognesi *et al.*, 2008). Its embryonic role is not known in butterflies, but *WntA*, along with the genes *optix* and *cortex*, is a major candidate gene for butterfly wing patterning which has been identified through decades of mapping studies in the *Heliconius* butterflies (Jiggins *et al.*, 2017). Association mapping studies have consistently implicated non coding changes around these loci in the evolution of wing patterning (Belleghem *et al.*, 2017; Hof *et al.*, 2016; Nadeau *et al.*, 2016; Wallbank *et al.*, 2016). Furthermore, *WntA* is expressed in the final instar larva of many species, and its expression pattern coincides with melanic elements in both forewing and hindwing regions (Martin and Reed, 2014; Martin *et al.*, 2012). In *Heliconius*, *WntA* expression specifically delineates the boundaries between light „bands“ and darker pigmented regions (Jiggins *et al.*, 2017; Martin *et al.*, 2012), while in species displaying classical nymphalid groundplans, *WntA* delineates the basal, central and border symmetry systems (See Chapter I - Figure 5; Martin and Reed, 2014). Additional evidence for the role of

*WntA* in wing patterning comes from heparin injection experiments. Heparin binds Wnt family ligands (Baeg *et al.*, 2001; Binari *et al.*, 1997; Fuerer *et al.*, 2010; Hufnagel *et al.*, 2006) promoting their mobility and resulting in phenotypes consistent with an expansion of *WntA* expression in adult wings (Martin and Reed, 2014). While these experiments make a compelling case, due to a lack of functional tools available for these species, no direct evidence for *WntA* in wing patterning is yet available. The CRISPR/Cas9 *WntA* protocol for *P. aegeria* was developed in collaboration with Arnaud Martin (University of Washington), who has developed such protocols for a number of other butterfly species (Mazo-Vargas *et al.*, 2017, see also Appendix IV - Additional File 1). The *WntA* homolog was recovered from a low coverage genome for *P. aegeria* (details regarding this genome can be found in Ferguson *et al.*, 2014). The same homologous region was used to design the sgRNAs as that for the other aforementioned butterflies by Mazo-Vargas *et al.*, (2017).

Butterfly wing patterns are laid down in a complex way, as they rely on a series of developmental events, each of which can affect the colour pattern observed (Brakefield, 2007; Jiggins *et al.*, 2017; Parchem *et al.*, 2007; Skelhorn *et al.*, 2016). First of all, the development of the wing *per se* affects colour patterning. Wing veins delineate the wing cells within which the pattern elements develop, and these affect the size, and colour, of the pattern elements. A complex interplay of both growth factors and patterning molecules play a role, with some being involved in both tasks (Weatherbee *et al.*, 1999; Keys *et al.*, 1999; Carroll *et al.*, 1994; Parchem *et al.*, 2007). Furthermore, *wingless* (*wg*) signalling is important in establishing these earlier wing patterning stages (Martin and Reed, 2010a). Therefore, while KO of *WntA* and *yellow* is likely to have an effect on the colour patterning and colouration of developing wings, larger scale effects visible as differences in wing shape and size might also arise. The effects on overall wing shape and size were investigated by means of geometric morphometrics. This technique allows accurate measurement of the positioning of homologous landmarks of wing veins within the wings (Breuker *et al.*, 2010). The relative shifts in position of these landmarks are then measured between treatment groups to assess the effect of CRISPR/Cas9 injections.

Since the primary focus of my PhD was unravelling the function and evolution of the *Shx* genes, I also designed sgRNAs against the *Hox3* paralogs in *P. aegeria*. The *Shx* genes are amongst some of the most widely variable homeodomain containing transcription factors described (Ferguson *et al.*, 2014 and Chapter II), and in stark contrast to other genes found within the *Hox* cluster. It is commonly accepted in the field of Evo-Devo that regulatory DNA evolution is the primary source of genetic diversity underlying phenotypic evolution (Carroll, 2000; Carroll, 2008; Hoekstra and Coyne, 2007), where morphological variation arises largely through *cis*-regulatory mutations affecting the expression of functionally conserved proteins. While such studies examining the impact of *cis*-regulatory changes to evolution abound, the contribution of gene duplication and divergence in the context of Evo-Devo remains relatively understudied. The variation exhibited by the *Shx* genes therefore provides an ideal system with which to study transcription factor evolution, both in the context of gene duplication between the paralogs, and between species.

In order to address this question, two *Shx* genes were initially targeted; *ShxC*, as it is important in the maternal specification of the serosa, and *ShxA* as it has a strong embryonic expression, and replaces the maternal expression of *ShxC*, with an overlapping zygotic pattern (See Chapter I and Figure I - 4). Given the localised expression of these genes, and their hypothesised role as extraembryonic tissue inducers (Ferguson *et al.*, 2014), one would expect serosal cells carrying the mutations to display changes in cell fates, thus resulting in visible developmental mutant phenotypes (Figure I – 4). However, screening of developmental defects for the *Shx* genes is complicated by the possible embryonic lethality and lack of a clear expected phenotype. Since only mosaic mutants are recovered from the injected individuals, screening for potential mutations in the serosa proved difficult, showing a limitation of this technique for early developmental genes. Furthermore, we observed a large increase in mortality for *ShxC* injected embryos. Previous results have shown some of the *Shx* genes are expressed as far as 48hAEL, including *ShxC* (Ferguson *et al.*, 2014), suggesting that they might be expressed under different developmental contexts. Moreover, a recent transcriptome recovered from the cabbage white, *Pieris rapae*, reported the presence of *ShxC* transcripts in larval tissue (Qi *et al.*, 2016).

Given the large sequence variability observed between orthologs in different species, and their fast evolutionary rates (Chapter II), it seems likely that the function of the *Shx* genes has diversified between different species, and may have assumed new (pleiotropic) roles during their evolution. In order to test whether *Shx* expression is maintained in the serosa in later stages, and to test for expression in other developmental contexts, *in situ* hybridisations were performed at later embryonic stages.

## **IV.2. Methods**

### **IV.2.a Animal Husbandry**

*Pararge aegeria* individuals were reared under laboratory conditions using lines derived from wild caught females from Belgium. The animals were reared on a 16-8h day-night cycle and kept at 23°C and 65% relative humidity. Larvae were fed on host plants, which include a mix of *Poa trivialis*, *Brachypodium sylvaticum* and *Dactylis glomerata*. Once eclosed, adult butterflies were fed from an artificial flower containing 10% honey solution (Gibbs *et al.*, 2004).

### **IV.2.b Cas9-mediated genome editing**

Target sequences were recovered from the draft *P. aegeria* genome, details of which can be found on Ferguson *et al.*, (2014) and accessed through lepbases.org (Challis *et al.*, 2016). Orthologous sequences from *B. mori* and, where available, other lepidopteran species were used to tBLASTN against the *P. aegeria* genome, and hits were extracted and manually annotated to produce gene models with intron exon boundaries. Phylogenetic trees were then created to check that each recovered sequence clustered with previously annotated orthologs from other species.

SgRNA target sites were designed by seeking sequences corresponding to N<sub>20</sub>NGG on exon regions of the sense or antisense strand of the target gene using the E-CRISP program (Heigwer *et al.*, 2014). Candidate target sequences were then BLASTed against the *P. aegeria*

genome to eliminate those with potential off-target sites using strict criteria, where the candidate editable site is defined only when the seed region (12 nucleotides (nt) to protospacer adjacent motif (PAM) NGG) is unique. From candidate editable sites, those with the first two bases of GG where available were selected, for sgRNA synthesis using a T7 promoter. The sgRNA templates were produced by PCR amplification with a forward primer encoding a T7 polymerase binding site and a sgRNA target site, and a reverse primer encoding the remainder of the sgRNA sequence (Bassett and Liu, 2014). The PCR templates were then purified using Qiagen PCR purification kit. sgRNAs were in vitro transcribed using the MegaShortScript T7 kit (Ambion). Cas9 protein was purchased from ThermoFisher (Catalog number: B25641). See Table 1 in Appendix IV for details on sgRNAs design and generation.

Injection mixes were prepared containing 400ng of each sgRNA and a final concentration of 333ng/ul cas9 protein. The mix also contained 0.05% phenol red which enabled monitoring the injection mix entering the embryos. Embryos were collected following 1 hour laying bouts from the host plants and aligned on a microscope slide using double sided sticky tape, with the anterior pole of the egg facing towards the outside. Microinjection of *P. aegeria* embryos was performed using a FemtoJet system with a pulled borosilicate glass needle (Harvard Apparatus, I.D.: 0.58mm). Injected embryos were then removed from the tape and transferred to Petri dishes lined with damp filter paper, placed in an incubator at 23°C and 65% relative humidity and allowed to hatch.

A proportion of embryos injected with the sgRNA targeting *ShxA* and *ShxC* were dechorionated and fixed at 8hr, 10hr, 16hr and 24hr after egg laying (hAEL) to determine the presence of possible serosal defects at those time points. To confirm DNA cleavage as a result of Cas9 activity we extracted DNA from adult legs (*yellow*) and embryos (*ShxA* and *ShxC*), PCR amplified the affected regions and looked for evidence of cleavage in agarose gels. Those gels which showed evidence of band separation were gel extracted and purified, and sequenced separately. The resulting traces were then aligned to check for evidence of cuts along the locus.

### **IV.2.c Imaging**

Following eclosion, animals were sacrificed by placing them overnight at -20°C. As soon as their wings had hardened, adults (both controls, and *WntA* or *y* injected) were mounted with their wings fully opened (Mazo-vargas *et al.*, 2017). The butterflies were photographed both dorsally and ventrally, using a Nikon D750 camera equipped with a Sigma 105mm f/2.8 lens and an off camera flash. The camera was mounted on an electronic slider, and used to take a picture at several position intervals to increase the depth of field, effectively producing a series of images that were subsequently z-stacked. Image files were imported into Adobe Lightroom 4.0, edited and exported to the program ZereneStacker 1.04, generating a fully stacked picture. Injected embryos were imaged using the AxioZoom V16 stereomicroscope (Zeiss).

### **IV.2.d Geometric Morphometrics**

The wings were subsequently removed from the thorax for more detailed morphometric analyses. The wings were flattened under a glass slide and photographed against a white background using a digital camera integrated with a stereomicroscope (Leica IC80 HD integrated into Leica MZ6; Las EZ software, version 1.8.0). Both left and right wings were photographed at 8x magnification. The order of photography was random with respect to sex and experimental group. A suite of 14 hindwing and 15 forewing landmarks (details to be found in Breuker *et al.*, 2007, 2010) were measured for each wing in ImageJ 1.6.0 (Schneider *et al.*, 2012). Measurements were done twice to accurately separate measurement error from biological shape variation in the analyses (i.e. Procrustes ANOVAs) (Klingenberg and McIntyre, 1998). Landmarks were locations where wing veins cross or where a wing vein meets the edge of the wing, and where chosen using the homology criterion (see Breuker *et al.*, 2010). The coordinates of the landmarks were used to calculate the centroid size, which is the square root of the sum of squared distances from a set of landmarks to their centroid (i.e. mean x and y coordinate of a set of landmarks per individual) (Klingenberg and McIntyre, 1998). Variation in wing shape in response to the *WntA* and *yellow* injections was investigated using geometric

morphometrics based on generalized least squares Procrustes superimposition methods (for details see Klingenberg and McIntyre, 1998). These analyses were done in MorphoJ 1.06d (Klingenberg, 2011). The Procrustes ANOVAs assessed shape variation between controls, *WntA* and *yellow* injected individuals, as well as whether significant asymmetry (i.e. difference in shape between left and right wing per individual) could be detected. The square root of the sum of the squared distances between corresponding landmarks of two optimally aligned configurations is an approximation of Procrustes Distance (Klingenberg and McIntyre, 1998). When the two optimally aligned configurations that are being compared are a left and a right wing of an individual, the Procrustes Distance is a measure of individual asymmetry of shape (Klingenberg and McIntyre, 1998). These Procrustes Distances were used to test whether CRISPR/Cas9 injections resulted in increased levels of asymmetry. Tests for differences in asymmetry were conducted using modified Levene's tests. This effectively meant fitting a linear model explaining variation in the left-right Procrustes Distances with size and experimental treatment as fixed factors. As the amount of shape variation can be size dependent (i.e. allometry), wing size was taken into account in the analyses (as log centroid size). Similar models were constructed to test for (log) size asymmetry differences as a result of the injections. Shape analyses were done two-fold: using only the 4 landmarks "inside" the wing (i.e. where wing veins cross), and those 11 landmarks that span the "outside" of the wing (i.e. where the wing veins meet the wing edges; cf. Breuker et al 2010) (Appendix IV – Figure 1).

A measure of wing size, not dependent on landmark positions, was determined for another set of analyses - overall wing size. Measurements corresponding to eyespot size and internal wing areas were performed three times manually in Image J 1.6.0, and an average taken, to compare wild-type individuals to CRISPR injected ones. The corresponding measurements were imported into R 3.3.3 (R core team, 2013) and two tailed student t-tests were performed to test for significant differences in measured areas as compared to wild-type animals. For those measurements that were not normally distributed, the non-parametric Mann-Whitney U test was performed (See Figure IV- 1 for details).



#### **IV.2.e *In Situ* Hybridisation of *Shx* genes**

In order to assess the potential roles of the Hox3 genes later in development, expression patterns were examined by means of *in situ* hybridisation. *In situ* hybridisation was performed on the 5 Hox3 genes (*Zen* and *ShxA-ShxD*), on a set of pooled embryos spanning ~0-120hAEL (reared at 23°C). The riboprobes used for the *in situ* hybridisation experiments are described in Ferguson *et al.*, (2014). For a detailed description of *in situ* hybridisation procedures in *P. aegeria* see Ferguson *et al.*, (2014) and Carter *et al.*, (2015). Briefly, a pool of embryos was collected from host plants, where females were allowed to oviposit for up to 120hrs. Collected eggs were dechorionated and fixed in 5% formaldehyde at 4°C overnight (cf. Carter PhD Thesis). Samples were then dehydrated in a methanol series (10%, 25%, 50%, 75%, 90%, and 100%), stored for up to one month, rehydrated and digested with proteinase K. Digested embryos were then exposed to a shorter round of fixation and washed extensively. Embryos were then incubated overnight at 55°C in Hybridisation solution (50% Deionised formamide, 5x SSC, 0.02% Tween 20, 100 µg/ml denatured Yeast tRNA, 2 mg/ml Glycine) containing 100 ng/µl of riboprobe. The embryos were then further washed and incubated in blocking solution (Roche Applied Science, Penzberg, Germany) for 30 min before anti-DIG antibody incubation at room temperature for 3–4 h. Excess antibody was subsequently removed through washes, and the staining developed using Alkaline Phosphatase buffer with NBT/BCIP. Embryos were subsequently imaged using the AxioZoom V16 stereomicroscope (Zeiss).

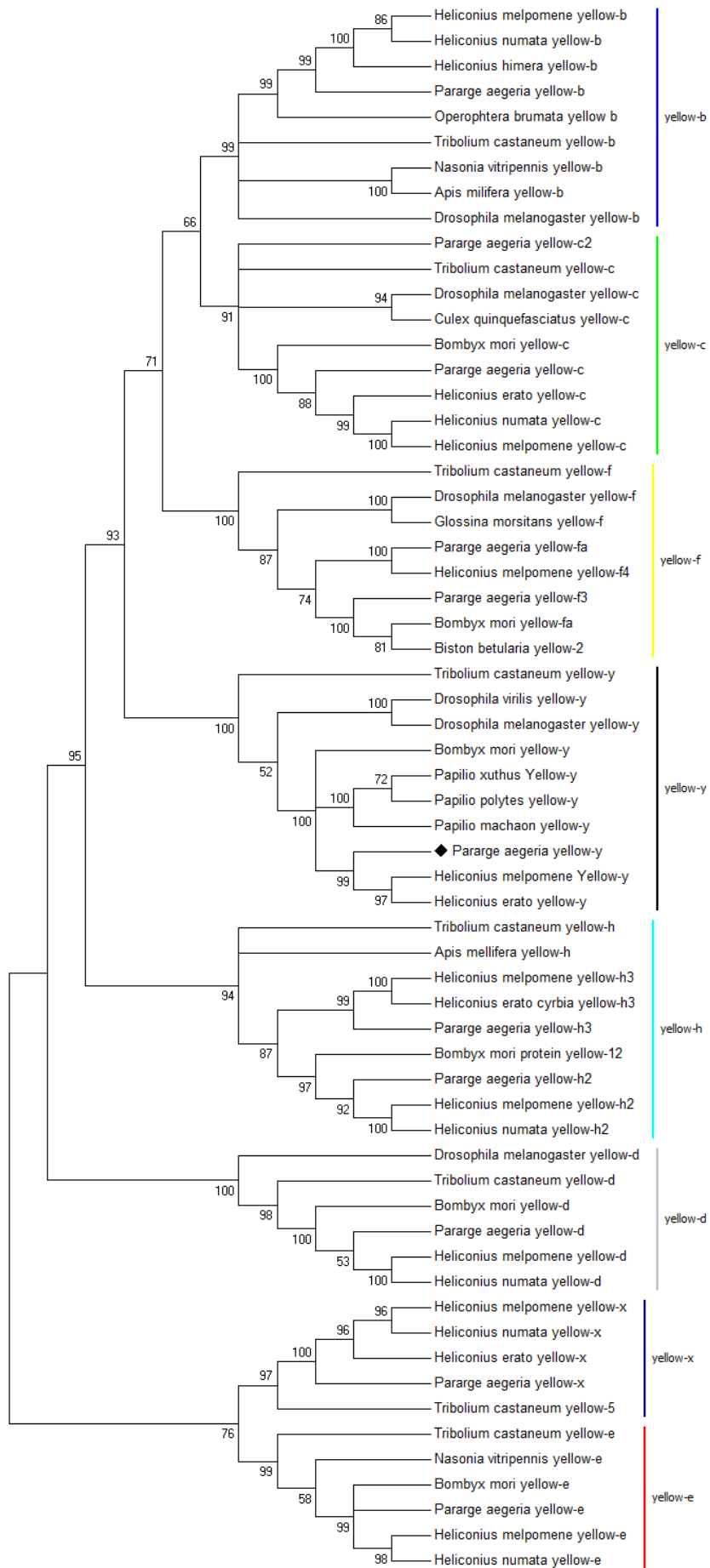
## IV.3. Results

### IV.3.a *yellow* knock-outs in *Pararge aegeria* wings

*yellow* encodes a secreted extracellular protein required for production of black melanin pigments in *Drosophila* (Biessmann, 1985; Wittkopp, True, *et al.*, 2002). Knockout of *yellow* in several species has been shown to result in loss of cuticular pigmentation (Wittkopp and Beldade, 2009). We therefore would expect a mosaic pattern of loss of pigmentation in the adult wings of *P. aegeria*. The phylogenetic tree produced from annotated orthologous sequences confirmed the identity of the extracted *P. aegeria yellow* sequence (Figure IV- 1). For phylogenetic analyses, sequences were recovered from data repositories where previously annotated orthologs have been deposited. The resulting trees are not representative of an exhaustive list of orthologous sequences, as they were designed to show the correct clustering of the identified *P. aegeria yellow* genes with other, previously described orthologs. For a detailed discussion of the insect *yellow* gene family see Ferguson *et al.*, (2011).

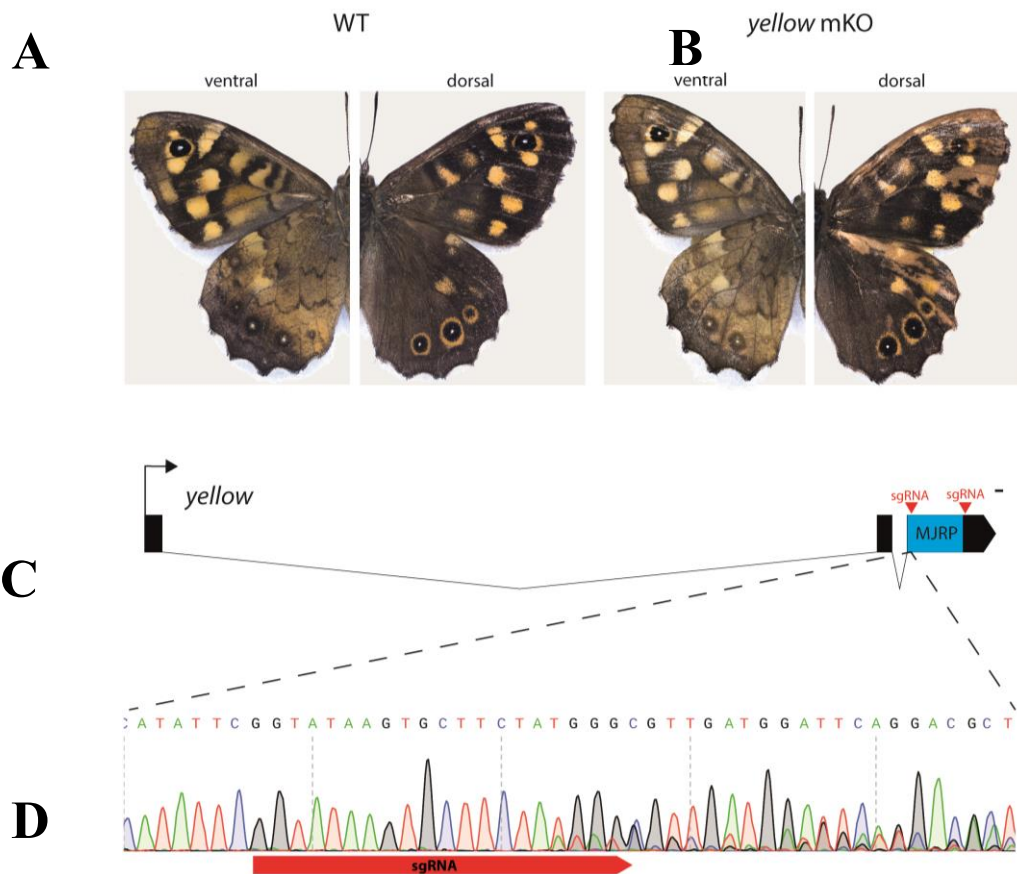
The aim was to produce *yellow* mutations through the introduction of a double stranded break by co-injecting two sgRNAs flanking the conserved major royal jelly protein (MRJP) motif (Figure IV- 2). The cut was confirmed by DNA extraction and Sanger sequencing, which showed only one sgRNA cut 3bp upstream of the NGG site. This was evident from the pool of mosaic cells as an indel signature present in the recovered sequencing traces (Figure IV- 2). Examination of injected individuals revealed ectopic loss of black pigmentation across both forewings and hindwings (Figure IV- 2), in accordance with previously reported results in other butterfly species (Perry *et al.*, 2016; Zhang *et al.*, 2017). The affected cells in the dorsal and ventral part of the wings were often asymmetrical, suggesting independent migration of affected cells during wing development. In the affected cells, areas of previous dark brown pigmentation became pale yellow, while black pigmented scales in the eyespots turned brown. Regions of yellow around the eyespots and in the forewings remained unaffected suggesting a crucial role for this gene in the synthesis of black melanin, while having no effect on lighter pigments of *P.*

*aegeria*. Interestingly, darker areas on the ventral side of the forewings seemed to be differently affected in terms of colouration as compared to the dorsal side, where the yellow around the hindwing eyespots is affected, while seeing no obvious effects on the yellow around the dorsal forewing eyespot (see also Appendix IV – Figure 3 for multiple mutant individuals). Loss of pigmentation in the abdomen and thorax of some individuals was also observed (data not shown).



**Figure IV- 1. Phylogenetic analysis of the *yellow* family genes in *Pararge aegeria*.**

The evolutionary history was inferred by using the Maximum Likelihood method based on the Le and Gascuel (2008) model. The model was chosen based on the automatic model prediction tool implemented in MEGA7 (Kumar *et al.*, 2016). The tree with the highest log likelihood is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 1.5516)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 1.1262% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 62 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. All branches with less than 50% bootstrap support are collapsed. There were a total of 283 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar *et al.*, 2016). The position of the *P. aegeria yellow* gene is highlighted with a rhombus in the tree.



**Figure IV- 2. Knock-out of the *yellow* locus in *Pararge aegeria*.**

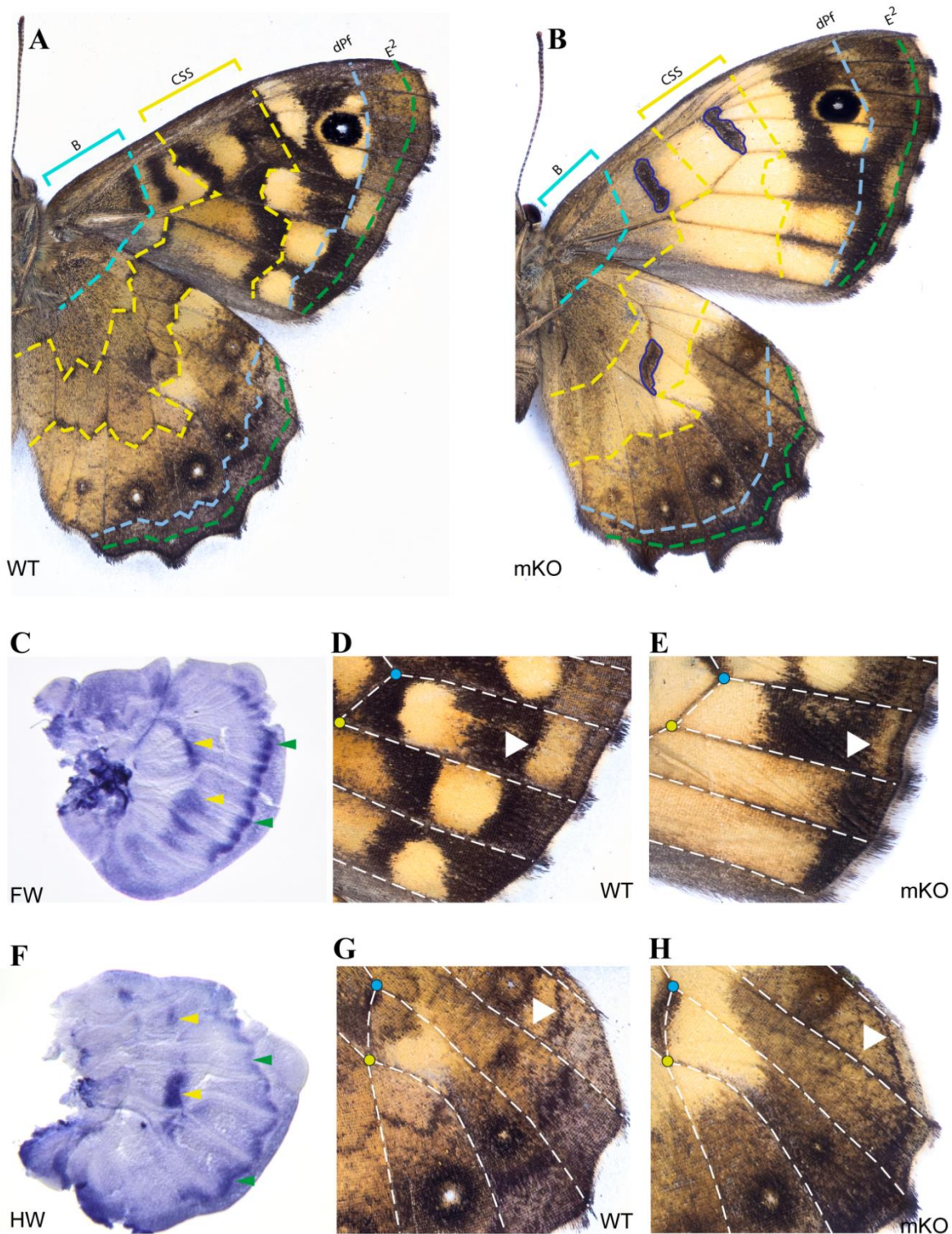
Wild-type *P. aegeria* showing ventral and dorsal wing morphology (A), shown alongside a representative *yellow* mKO individual in (B). The sgRNAs were designed to target the MJRP motif shown in the schematic of the *yellow* gene (C). Sanger sequencing traces of pooled individuals confirmed cut (D).

### IV.3.b *WntA* Knock-outs in *Pararge aegeria* wings

*WntA* is a member of the highly conserved Wnt signalling ligands. Expression of *WntA* in butterflies coincides with patterns in the basal, central and marginal symmetry systems of butterfly wings (Gallant *et al.*, 2014; Gallant *et al.*, 2014; Martin and Reed, 2014), and has been proposed to act as a morphogen, sitting on top of the pigmentation cascade and dictating areas of melanisation in a concentration dependent manner (Jiggins *et al.*, 2017). The extracted *WntA* sequence from the *P. aegeria* genome clustered with other previously annotated *WntA* orthologs, confirming the identity of the sequence (Appendix IV, Figure 4). We designed two sgRNAs targeting exons 4 and 5 of the gene (Mazo-vargas *et al.*, 2017), and were able to induce mutations in several injected individuals. A total of 830 embryos were injected over the course of three days, where 207 hatched (24.9%). We were able to recover 32 individuals with mosaic phenotypes, giving 12% efficiency in the rate of mutation in the F<sub>0</sub>.

Mosaic mutant individuals show an expanded area of yellow pigmentation as compared to the wild-type animals corresponding to defects in the basal, central and marginal symmetry systems of the *P. aegeria* wings (Figure IV- 3, A-B). In accordance with expression patterns described in other species (Martin and Reed, 2014), the discal systems remain unaffected, as these are associated with expression of other Wnt ligands. In individuals carrying a large proportion of mutant cells, most of the dark pigmented cells in the basal and central symmetry systems become pale yellow, and most patterning respective to wild-type yellow/dark boundaries in pigmentation are abolished. In the marginal system, mosaic individuals display a reduction in the distance between the distal parafoveal (dPf) element and the wing margin, as well as a reduction in the yellow patch bordering the dPf element and the external systems ( $p < 0.001$ , Figure IV- 4). These results are in accordance with expression of *WntA* in *P. aegeria*, which localises to these three systems in 5<sup>th</sup> instar wing discs (Mazo-Vargas *et al.*, 2017, Figure IV- 3, C and F). Like the *yellow* mKOs, the mosaic patterns of mutations are independent in the ventral and dorsal sides within the same individual, suggesting asymmetric migration of scale cells between dorsal and ventral sides of the wings (McMillan *et al.*, 2002). Some affected

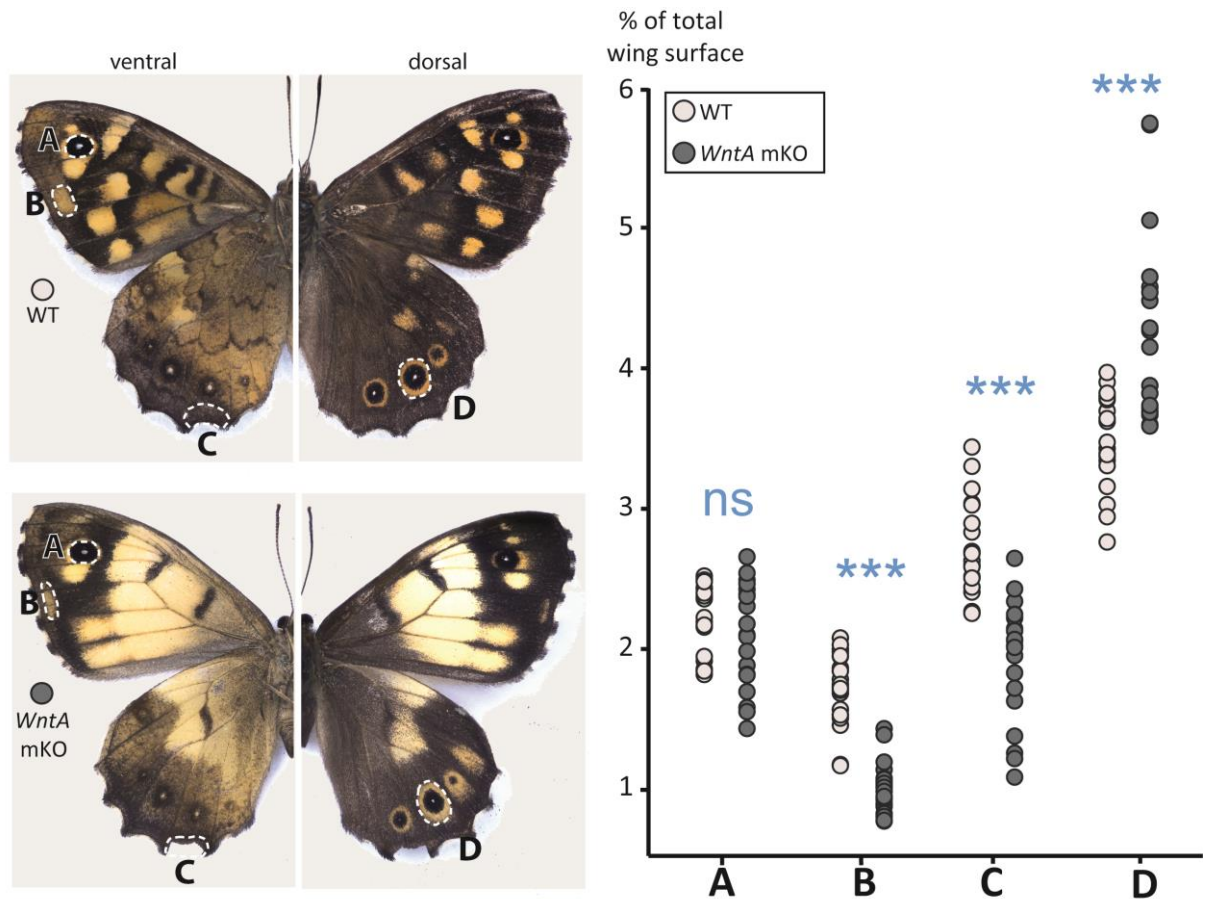
individuals also show a large reduction in eye-spot size and colouration (See Appendix IV, Figure 2). Size reduction is evident in both the dorsal and ventral sides of the wings and the area of white pigmentation within the eyespot is almost completely abolished in some individuals. The area of yellow pigmentation around the eyespots is also affected, becoming enlarged and taking on a paler colouration with respect to the wild-type eyespots. In one individual, there was a complete loss of one eyespot in the hindwing. The loss was symmetrical and occurred on both left and right hindwings (See Appendix IV, Figure 2 for multiple mutant individuals). However, there was no statistical significance associated with fore wing eyespot size in Wild-type and *WntA* mosaic individuals ( $p = 0.233$ , Figure IV - 4). On the other hand, there was a significant difference between *WntA* mutants and WT individuals with respect to the yellow pigmented area around the hindwing eyespots ( $p < 0.001$ , Figure IV- 4).



**Figure IV- 3. *WntA* Knock-out in *Pararge aegeria*.**

Wild-type individual (A) showing the basal (B), central (CSS) and marginal (dpf, E<sup>2</sup>) symmetry systems of *P. aegeria*. *WntA* KO abolishes these systems in affected individuals (B), while the discal elements remain unaffected (blue outlines). *WntA* expression in 5<sup>th</sup> instar larval imaginal discs coincides with the central (yellow arrow heads) and marginal (green arrow heads) symmetry systems in both forewings (C) and hindwings (F). Marginal systems are distalised in both forewings (D-E) and hindwings (G-H). The distalisation of the dpf is indicated with the white arrow head in both.



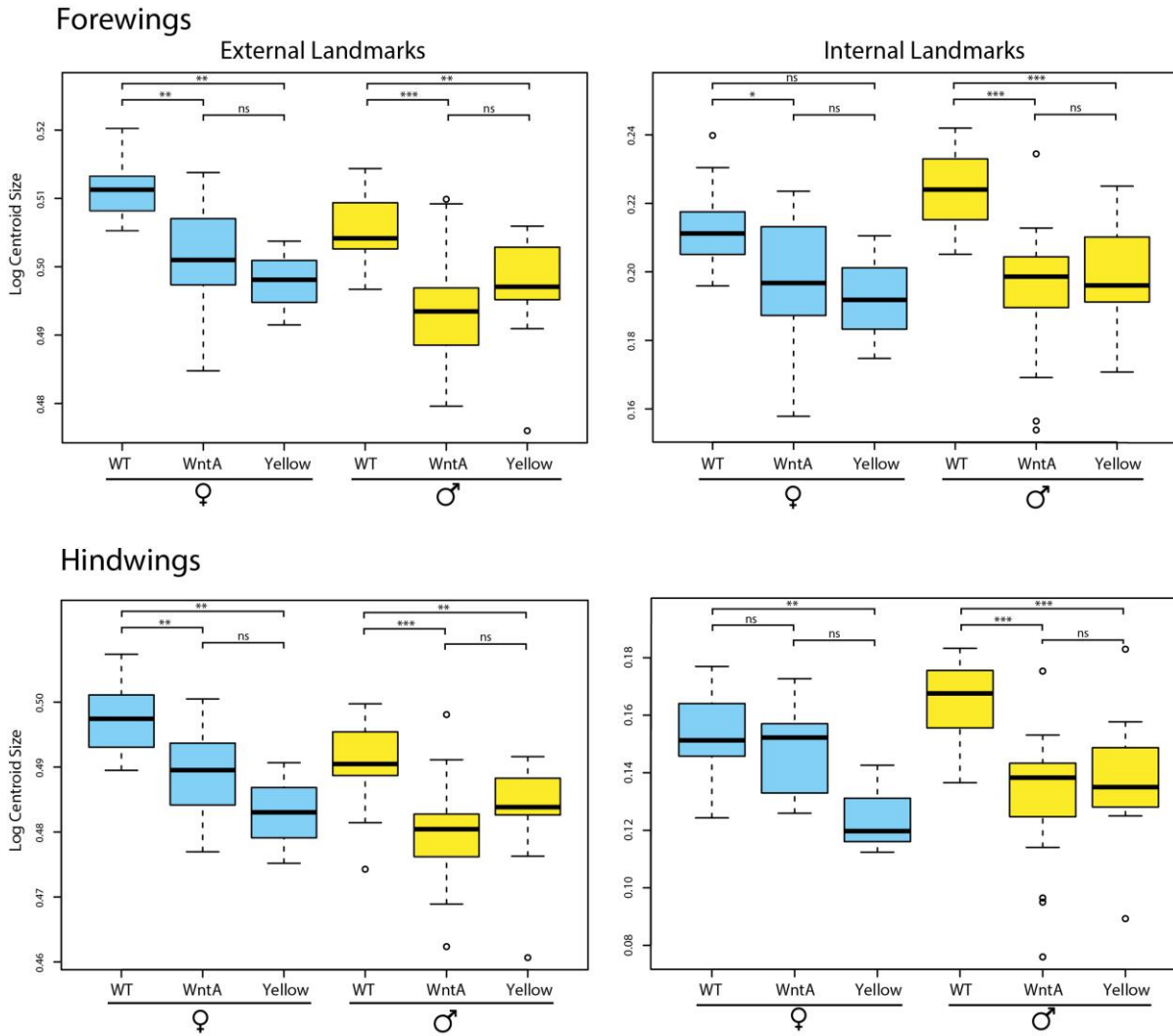


**Figure IV- 4. Effects of *WntA* mKO on marginal elements of *Pararge aegeria*.** (A) No effect of *WntA* loss-of-function in forewing eyespot rings. (B) Distalisation of the ventral forewing M2-M3 dPF element, resulting in a reduction of the light colour-area ( $p < 0.001$ ). (C) Distalization of the ventral hindwing M1-M2 dPf element ( $p < 0.001$ ). (D) *WntA* loss-of-function results in expanded dorsal hindwing eyespot outer rings ( $p < 0.001$ ).

### IV.3.c *WntA* and *yellow* effects on shape and size of *Pararge aegeria* wings – Geometric Morphometrics

While KO of both *yellow* and *WntA* in *P. aegeria* have been shown to result in pigmentation and patterning defects, it is also possible that they may affect wing size and shape (Johnson *et al.*, 2014). A geometric morphometrics approach to test the contribution of injections to overall shape and size revealed that CRISPR/Cas9 injections had a significant effect in all cases for overall size (Appendix IV– Table 2). That is, both internal and external landmarks for both females and males were differentially affected as a result of the injection. When looking for which specific treatment had an effect on overall size, both the *WntA* and

*yellow* injected individuals showed an overall significant decrease in wing size compared to controls, but there were no differences in wing size between *WntA* and *yellow* (Figure IV – 5). There was also an effect on wing size asymmetry but only for female hindwing internal landmarks (Appendix IV– Table 2).



**Figure IV – 5. Effects of CRISPR/Cas9 injections on overall wing size in *Pararge aegeria*.** Injections result in an overall decrease of wing size between wild-type individuals and injected groups for both forewings (top) and hindwings (bottom). External (left) and internal (right) landmarks for both males and females are shown separately. Significant effects were not reported in any case between *WntA* and *yellow* injected animals. \* significant at  $p < 0.05$ ; \*\* significant at  $p < 0.01$ ; \*\*\* significant at  $p < 0.001$ .

The results also show that there was a significant effect on shape difference as a result of treatment on male forewing and hindwing external landmarks, as well as internal landmarks for both forewings and hindwings (Appendix IV – Figure 4 and Appendix IV– Table 2). There was also a significant effect on wing asymmetry for forewing internal landmarks in both males and females, as well as external landmarks for males only in forewings and hindwings (Appendix IV – Figure 4).

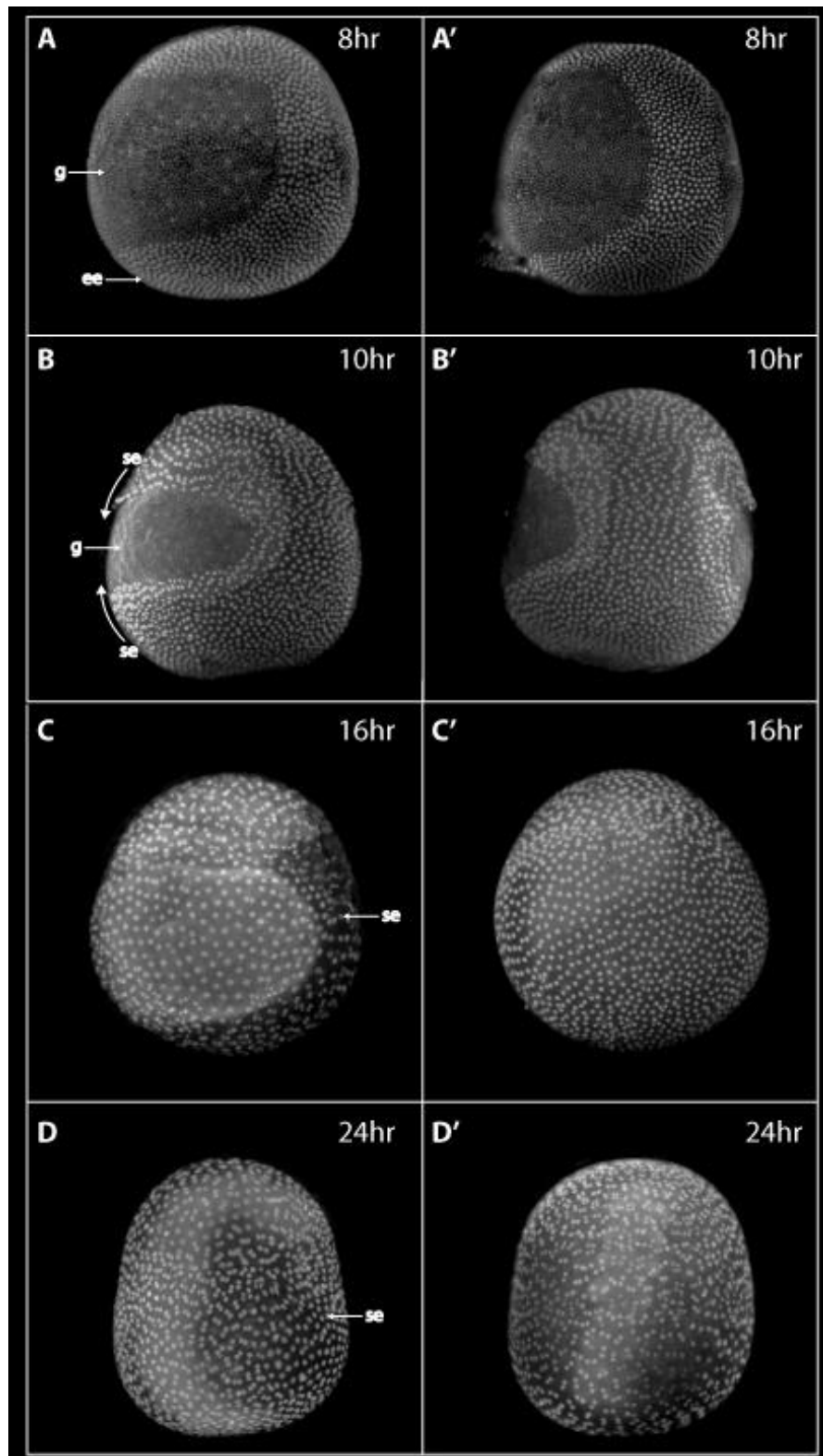
#### **IV.3.d Knock-out of *ShxA* and *ShxC* in *Pararge aegeria***

For both *ShxA* and *ShxC*, two sgRNAs were injected, aimed to disrupt the likely functional homeodomain of both genes (See Chapter II; Figure IV- 7). Injection of both *Shx* sgRNAs resulted in a large increase in mortality rates as compared with other injections, where only 4 individuals hatched (personal observation, not recorded). Only a single adult survived to eclose from these injected animals, and the phenotype of this individual did not significantly differ from the control phenotype. DNA extraction from a pool of injected embryos confirmed cuts did occur at both loci however (Figure IV- 7). For *ShxA*, we recovered Sanger sequencing traces which showed a ~200bp deletion between the two injected sgRNAs, completely deleting part of the second exon encoding the homeodomain. Sequencing of *ShxC* revealed an ~85bp deletion spanning the homeodomain (Figure IV- 2).

To determine serosal phenotypes, embryos were observed following fixation at 8hAEL, 10hAEL, 16hAEL and 24hAEL. Between 8 and 12hAEL, *ShxA* is strongly expressed in the serosa (Ferguson et al., 2014), and serosal closure is completed at around 12hAEL. However, no clear serosal phenotypes were observed at any of these stages, with injected embryos having wild-type like serosal cells (Figure IV- 6). At 8hAEL, cells corresponding to extraembryonic tissue have started differentiating, and appear larger and polyploid as compared to germ band cells. By 10hAEL, serosal cells start migrating over the germ band, which becomes encapsulated by around 12hAEL. While several undeveloped embryos were observed in the injected batch (probably due to death upon injection), the embryos that did develop displayed

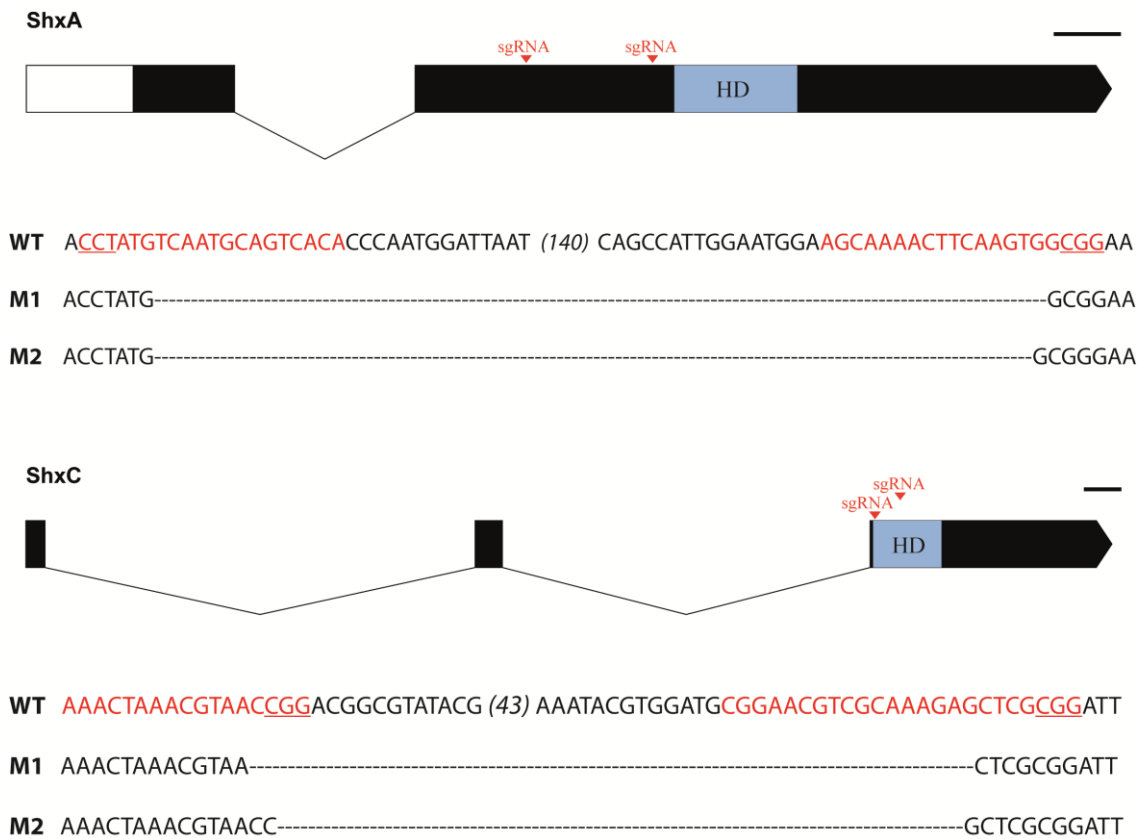
the same cell structure and morphogenetic movements as wild-type embryos (representative individuals shown in Figure IV- 6, A'-D').

Of the pool of embryos that were fixed at each time point, there were a large percentage of embryos showing no signs of development. Pooling the time points together, on average, 18.2% of embryos had visible signs of development for *ShxA*, and 17.3% for *ShxC*. This could be a direct result of the injections *per se*, but it is interesting to note the higher mortality associated with these genes compared to the *WntA* injections.



**Figure IV- 6. CRISPR/Cas9 injections for *Shx* in *Pararge aegeria*.**

DAPI stainings of wild-type embryos at 8, 10, 16 and 24hAEL are shown on the left hand side (A-D). At 8hAEL, germ band (g) and extraembryonic cells (ee) have started differentiating. By 10hAEL, the serosal cells start migrating and encapsulate the germ band by about 12hAEL. The serosal cells are present as large polyploidy cells fully encapsulating the developing embryo at 16 and 24 hAEL. Injected embryos show normal signs of development (A'-D').



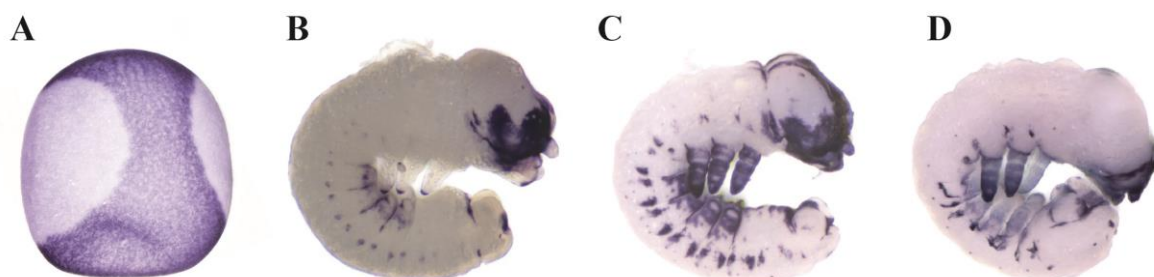
**Figure IV- 7. Targeted locations of the *ShxA* and *ShxC* loci in *Pararge aegeria*.**

Gene maps and respective sgRNA locations for *ShxA* and *ShxC* are shown. The wild-type sequence is shown immediately beneath the gene maps, with sgRNAs highlighted in red and PAM sequences underlined. Two mutant traces were recovered per gene, shown below as M1 and M2. The number in brackets indicates the number of nucleotides between the alignments shown. Exons shown as black boxes, introns as lines and UTRs as white boxes. HD = Homeodomain, black bar = 100bp.

### IV.3.e *ShxC* expression in embryonic tissue

*In situ* hybridization revealed no detectable expression of any of the *Hox3* genes in serosal tissue after around 15hAEL. Expression of *ShxC* was however detected in embryos ~70 to ~90hAEL (Figure IV - 8). Presence of *ShxC* transcripts around this time point is confirmed by transcriptomic data (unpublished), with other *Shx* genes also being present, but at very low levels. Serosal expression of *ShxC* was confirmed (Figure IV - 8, **a**), with a complex and dynamic expression pattern observed in embryonic tissue much later in development. At around 70hAEL, *ShxC* transcripts are localised to the distal end of the developing limb buds and prolegs. Expression is also observed at the base of the invagination giving rise to most

segments, and at the telson. Strong expression is observed in the developing head capsule. At ~80hAEL expression domains expand proximally in developing limb buds and prolegs, where ring-like expression is observed around their developing segments. Strong expression at the base of the developing segments remain, as well as in the head capsule. At ~90hAEL, similar expression domains are maintained, while additional expression is localised to the spiracles.



**Figure IV – 8. *ShxC* expression in *Pararge aegeria* embryos.**

*ShxC* expression localises to extraembryonic tissue at ~8hAEL (A). Expression of *ShxC* is then observed in the embryonic tissue at ~70hAEL, localizing to the distal end of the developing limb buds and prolegs (B). Expression is also observed in the head capsule, as well as at the base of the invaginating segments. At ~80hAEL, *ShxC* expression expands in the developing limb buds and prolegs, forming ring-like domains that strongly localise to segment boundaries. Expression in the head capsule is retained, as well as at the base of the segments, which are now more pronounced (C). Ring-like expression in developing limb buds is maintained at ~90hAEL, with expression subsiding in other domains (D). Localised regions of expression become pronounced in the developing spiracles.

#### IV.4. Discussion

The development of functional tools for the study of gene function is crucial to gain an understanding of the role genes play in development, and to better understand how they might evolve to give rise to phenotypic differences between and within species. This chapter outlines the recent progress made in developing functional tools in the Lepidoptera, and in particular the successful application of the CRISPR/Cas9 system in my model species *P. aegeria*.

#### IV.4.a *Yellow* has a conserved pigmentation role in *Pararge aegeria*

Through the microinjection of a sgRNA complementary to the *yellow* CDS, mutations were introduced that resulted in cells carrying knock-out (KO) phenotypes for the gene. In accordance with recently published results in other butterflies (Perry *et al.*, 2016; Zhang *et al.*, 2017), KO of *yellow* results in a loss of melanisation on the wings of *P. aegeria*. These results are also in accordance with expectations from phenotypes observed in *Drosophila* (Wittkopp, True, *et al.*, 2002), as well as other insects including *Tribolium castaneum* (Arakane *et al.*, 2010), *Oncopeltus fasciatus* (Liu *et al.*, 2016) and *Bombyx mori* (Futahashi *et al.*, 2008). The *yellow* KO results in *P. aegeria* thus support a conserved role for this gene being involved in melanin pigmentation. It is interesting to note that the KO phenotypes were observed in both hindwings and forewings, as well as in the dorsal and ventral sides of the wings. This is similar to the findings reported in *Drosophila*, but different to the RNAi phenotypes observed in *Tribolium* and *Oncopeltus*, where only hindwing defects were noted (Arakane *et al.*, 2010; Liu *et al.*, 2016). This discrepancy is also in accordance with a recently published study, where several species of butterflies with *yellow* mutants also contained both forewing and hindwing defects (Zhang *et al.*, 2017). These differences suggest that deployment of *yellow* has diverged throughout insect evolution.

In several *Drosophila* species, variation in abdominal and wing pigmentation has been linked to changes in *cis*-regulatory elements controlling the expression of *yellow* (Gompel *et al.*, 2005; Werner *et al.*, 2010). For example, the evolution of the wing spot of *D. biarmipes* has been linked to the evolution of a *cis*-regulatory element which ancestrally functioned to give a low level of expression across the whole wing (Arnoult *et al.*, 2013). Following the co-option of transcription binding sites for Engrailed and Distalless, *yellow* became associated with expression in the epidermis prefiguring the wing spot. In another classical example of gene regulatory network co-option, the gene *wingless* (*wg*) has been linked to the evolution of melanic spots observed on the wings of *D. guttifera*, by the acquisition of binding sites for *wg* at the *yellow* locus (Werner *et al.*, 2010).



Given the highly conserved role of *yellow* throughout insect evolution, and its repeated association with the evolution of wing pigmentation, it is a possibility that changes in melanic regions of *P.aegeria* have evolved through differences in *cis*-regulation of the locus. This could in part provide an explanation for the large variation observed in melanic forms along the North-South cline of Europe.

#### **IV.4.b Evolutionary ecology of *Pararge aegeria* wing melanisation**

The *WntA* results in particular (Figure IV – 2) indicate that there is one area on the *P. aegeria* wings that remains largely melanised whilst other areas become less melanised after knock-out; the dorsal basal forewing area. Note, not the corresponding ventral area. Thus, one may conclude that, that part of the wing relies on somewhat different pigmentation mechanisms. From an evolutionary ecology perspective this is potentially an interesting finding. Many butterflies, including *P. aegeria*, rely on overall dorsal wing melanisation, but this basal part of the wing in particular, to heat up in order to become active (Talloe *et al.*, 2004; Berwaerts *et al.*, 2001). Not only does there appear to be genetic variation between populations associated with melanisation (i.e. populations from Northern Europe being darker), but there is also a large degree of developmental plasticity in melanisation in response to temperature and diet (e.g Nylin *et al.*, 1995), not least because the production of melanin is costly and subject to various resource trade-offs. Interestingly enough, despite clear effects on wing development (in particular asymmetry; *in prep*), no significant effect of baculoviral infection on *P. aegeria* wing melanisation has been detected (Hesketh *et al.*, 2012). In the study by Hesketh *et al.* (2012) there was an increased duration of development time, which may have compensated for any resource trade-off. One study concentrated on the size of the dorsal yellow areas in the border and central symmetry system, rather than the overall degree of melanisation. The size of the dorsal yellow areas has been associated with a degree of crypsis and thermoregulation, and associated different flight behaviours (Dyck and Matthysen, 1998). Males with less melanised wings and larger yellow patches spend significantly more time basking in sunlit patches, and darker males fly more frequently, suggesting an adaptive advantage to melanisation with respect to

thermoregulation and patrolling behaviour (Van Dyck *et al.*, 1997). The developmental precision with which these yellow patches are produced has a low, but significant, heritability in *P. aegeria*, with some yellow patches in the border symmetry system showing developmental integration (Van Dongen and Talloen, 2007; for developmental integration in border symmetry system see (Breuker *et al.* 2007).

It is therefore possible that changes in the regulation of *yellow* as well as *WntA* and or upstream factors have in part led to the difference in pigmentation observed between different populations of *P. aegeria*. Furthermore, the different symmetry systems appear to have a degree of developmental modularity, and thus different evolutionary potential, and are able to respond differently to selection.

#### **IV.4.c *WntA* has a conserved role in wing patterning in butterflies**

Decades of mapping studies in *Heliconius* butterflies have identified three major wing patterning loci associated with differences in colour variation (Jiggins *et al.*, 2017; Merrill *et al.*, 2015). One of these is the transcription factor *optix*, responsible for the red colouration often observed as bands in the forewings and stripes in the hindwings of several *Heliconius* species (Reed *et al.*, 2011). Another, more recently identified, is the member of the *fizzy* family of cell cycle regulators, *cortex* (Hof *et al.*, 2016; Nadeau *et al.*, 2016). This gene has evolved rapidly during the radiation of the Lepidoptera and is thought to play a role in wing patterning by controlling heterochrony of scale development (Koch *et al.*, 2000; Nadeau *et al.*, 2016; Jiggins *et al.*, 2017). Finally, *WntA*, a member of the highly conserved Wnt signalling pathway has been associated with various aspects of wing patterning and colouration (Papa *et al.*, 2013). Mapping studies for all three loci have consistently implicated non-coding, regulatory changes in the evolution of these genes, rather than coding variants (Belleghem *et al.*, 2017; Gallant *et al.*, 2014; Hof *et al.*, 2016; Wallbank *et al.*, 2016).

*WntA* is expressed in the final instar larvae of several butterfly species (Gallant *et al.*, 2014; Martin and Reed, 2014; Martin *et al.*, 2012) where it shows a large diversity in

localisation, and prefigures regions of dark pigmentation. Furthermore, heparin injection experiments, which promotes the mobility of Wnt ligands, lead to an increase in melanic pigmentation associated with wings (Gallant *et al.*, 2014; Martin and Reed, 2014). The combination of the genetic mapping and morphogen experiments provide a strong case for the role of *WntA* in wing patterning, but until now no direct functional evidence was available for any butterfly species. Knock-out of *WntA* in *P. aegeria* has shown that clones carrying the mutations for the locus result in expanded areas of light pigmentation, a result consistent with the predictions from *in situ* hybridisation and heparin experiments in other species.

In contrast to the role of *optix*, which has only been mapped in relation to the *Heliconius* butterflies, *WntA* expression is associated with adaptive pattern variation in a larger range of butterfly species (Gallant *et al.*, 2014; Martin and Reed, 2014). Interestingly, mapping studies in both *Heliconius* and *Limenitis* butterflies, have shown that a homologous non coding region upstream of *WntA* is associated with melanic pigmentation in both species groups, which diverged around 65Ma (Bellegheem *et al.*, 2017; Gallant *et al.*, 2014). This strongly suggests repeatability across large phylogenetic scales with regards to the evolution of wing patterning.

With this in mind, the *P. aegeria* injections were performed as part of a larger study examining the role of *WntA* in 6 other butterfly species; *Danaus plexippus*, *Vanessa cardui*, *Junonia coenia*, *Agraulis vanillae*, *Heliconius sara* and *Heliconius erato demophoon* (Mazovargas *et al.*, 2017). This study provides the first large scale phylogenetic functional study examining the role of adaptive wing colour variation in Lepidoptera, with species separated by a maximum of 90Ma of evolution (Wahlberg *et al.*, 2009). The availability of knock-out phenotypes for this diverse group of species allows for a comparison of the role of *WntA* in the more basal, classic Nymphalid ground plan species, with the more derived Heliconiinae which display a much more divergent mode of wing patterning.

Nijhout (Nijhout, 1978) famously proposed a shared ground plan composed of repeated pattern elements referred to as symmetry systems which have been hypothesised to arise from morphogenetic sources in the developing wing (Nijhout *et al.*, 2003; Otaki, 2011). *In situ*

hybridisation results across a wide range of butterfly species has shown that *WntA* marks the development of basal, central and external symmetry systems, while other Wnt ligands (*wg/Wnt6/Wnt10*) are involved in delineating the discal and external systems (Gallant *et al.*, 2014; Martin and Reed, 2014; Martin *et al.*, 2012). These results fit the KO phenotypes observed in *P. aegeria*, where patterning is abolished with respect to the basal and central symmetry systems (CSS), also where *P. aegeria WntA* is expressed during 5<sup>th</sup> instar wing development (Figure IV- 3, C, F). Interestingly, areas of dark pigmentation corresponding to the putative discal elements in both fore- and hindwings remain unaffected, as these are associated with expression of *wg*, rather than *WntA* in other species (See figure 2 in Appendix IV for multiple mutant individuals). This observation is consistent with phenotypes recovered from both *J. coenia* and *V. cardui*, where the basal and CSS systems are also affected, while always retaining discal elements (Mazo-Vargas *et al.*, 2017). This effect is striking in *J. coenia*, where the basal system disappears, and the two discal systems remain completely unaffected, despite having the exact same colour composition. This highlights the malleability of different Wnt ligands to produce similar colour wing patterns across the wing surface. This apparent genetic decoupling across symmetry systems in butterfly wings likely allows for accelerated rates of phenotypic evolution, where the precise spatiotemporal expression of different inducers can affect specific patterns without changing other symmetry systems within the developing wing.

*WntA* in *P. aegeria* and the other nymphalids is also expressed in the marginal section of the wings, and *WntA* KO in these regions results in a contraction of the marginal symmetry system as well as in a shift in the distal parafocal elements bordering the wing (Figure IV- 3). *WntA* is thus also deployed in the marginal wing elements, where its modulation likely contributes to changes in patterning dynamics at the wing border. It is also interesting to note that, in contrast to the other nymphalids, *WntA* expression is also localised to the forewing eyespots in *V. cardui* (Martin and Reed, 2014), and the eyespots are affected by *WntA* KO in this species. It is therefore likely that *WntA* was recently co-opted in the eyespot gene regulatory network in the lineage leading to *V. cardui* where it further elaborates upon the function of this complex structure.

While *WntA* KO consistently results in the abolishment of dark pigmentation along the basal and CSS system in both *P. aegeria* and *J. coenia*, *V. cardui* shows conflicting results between fore and hindwings. Loss of patterning along the CSS in hindwings is also observed as a loss of dark pigmentation in *V. cardui*. However, *WntA* KO results in expanded areas of dark pigmentation in the forewing. This apparent dual role as pattern inducer and pattern inhibitor is also observed in the more derived *A. vanillae*, where *WntA* KO results in both expansion and abolishment of white pattern elements, depending on their position in wing compartments. In *Heliconius*, *WntA* appears to play a simpler role, where precise expression of *WntA* is required for the delineation of the white/red bands in *H. sara* and *H. erato* respectively. *WntA* knockout thus results in expansion of both red and white bands in these species. Interestingly, the expansion of these elements is reminiscent of the wing colouration of several sub species of *H. sara*, suggesting that the *cis* regulation of *WntA* might partially underlie differences in this species complex.

Overall, these results argue for a conserved role of *WntA* across butterfly evolution, where repeated re-deployment of *WntA*, likely due to changes in regulatory elements, allows for the modulation of symmetry systems across the nymphalidae.

#### **IV.4.d CRISPR/Cas9 injections may have an effect on overall wing size and shape**

Geometric morphometric analyses revealed that overall wing size was reduced as a result of injections. However, due to a lack of a sham injected control group, it is difficult to conclude whether the reduction in size was due to the injection *per se*, or as a result of the mutations at the targeted loci. Indeed, wing size has been shown to vary as a result of environmental stressors in several species of butterflies (Johnson *et al.*, 2014; Pellegrons *et al.*, 2009). Moreover, the fact that both the *WntA* and *yellow* injected groups show the same direction in size reduction (they are not significantly different from each other), suggests that both groups responded similarly to the injections, irrespective of CRISPR/Cas9 treatment. Differences in overall shape were also recorded as a result of injection, as well as significant

instances of asymmetry within individuals (Appendix IV – Figure 4 and Appendix IV – Table 2). In terms of injections affecting overall shape, there were significant differences in both landmarks for both forewings and hindwings, but only for males. It is possible that the low sample size for female wings resulted in a lack of power to detect these differences, rather than it being a result of a dimorphic effect of the knock-outs. Significant asymmetry between left and right wings was also reported, for both sexes for forewings in internal landmarks only, but only males for both wings in external landmarks. Since injections are expected to result in mosaic mutants, the asymmetry might be a result of shape changes in wings carrying a larger proportion of mutant cells within individuals. The migration of cells is likely to be random with respect to the initial mutant cells, and the discrepancy in terms of affected wings and landmarks could be a result of this asymmetric migration of cells following the injections. However, as for wing size, it is difficult to conclude whether shape differences are a result of injection *per se*, or the effect of the mutated loci (Johnson *et al.*, 2014).

The fact that *WntA* injections might not affect wing shape is an interesting observation. Given that some mKOs appear to span the entire wing surface, this would suggest that *WntA* is not involved in wing morphology and its function might be limited to directing the expression of downstream genes involved in pigmentation. This is in contrast to other Wnt ligands, such as the canonical *wingless* gene, which is involved in patterning both the wing morphology (Carroll *et al.*, 1994) as well as directing the patterning of pigmentation in other symmetry systems within the wing (Martin and Reed, 2010a; Martin and Reed, 2014). Transgenic *WntA* mutants have now also been reported in *H. sara* (Jiggins, unpublished), showing that full KOs are able to survive in the F1, without any obvious effects on overall wing morphology. Given that *WntA* is known to be expressed in a segmental pattern during embryogenesis in butterflies (Arnaud, pers. comm.), this is a remarkable result, and suggests that Wnt ligands might be partially redundant, and able to escape consequences of pleiotropy in a way other than through *cis* regulatory modulation. Redundancy in Wnt ligands has indeed been previously described in *T. castaneum* (Bolognesi *et al.*, 2008).

#### IV.4.e *Shx* genes are likely pleiotropic

Injections targeting the homeodomains of *ShxA* and *ShxC* showed successful deletion of part of the homeodomain through Sanger sequencing (Figure IV- 7). Although a higher rate of mortality was observed compared to other injections, no serosal phenotypes were observed. However, embryos at later stages of development were unfortunately not screened, and the possibility of embryonic mutant phenotypes at later stages cannot be ruled out. Given that there is evidence of *Shx* expression at later stages in development (Figure IV -8), it is likely that the resulting mortality is induced by other pleiotropic effects other than serosal development.

The use of CRISPR/Cas9 for the study of gene function in early embryos, and on the basis of the results presented here also in extraembryonic tissue, is limited for several reasons. The resulting mosaicism from injections in the G<sub>0</sub> confounds the interpretation of mutant phenotypes, as only partial KOs are expected. Furthermore, as is the case for the *Hox3* genes in Lepidoptera, overlapping expression patterns of the various paralogs may suggest a possible redundancy in function, where mutant phenotypes might only be observed when KOs are performed in combination targeting multiple paralogs in parallel. The large increase in mortality is intriguing however, and could potentially be a result of mutations affecting later developmental stages. *ShxC* is indeed expressed later in development under a different developmental context (Figure IV - 8), and the very low hatching rate might have been a result of mutations affecting these later stages. The expression patterns for these later stages became available after the CRISPR experiment was performed, and unfortunately phenotypic defects at these stages were not recorded. Future injections aimed at the *Shx* genes should be performed with this in mind and allowed to develop fully, before screening for developmental defects.

The complex expression patterns retrieved for *ShxC* at later developmental stages are intriguing. An embryonic function for *zen* has been reported in *Drosophila*, where it is required to repress genes involved in the establishment of the optic field (Chang *et al.*, 2001).

Localisation of *ShxC* transcript was observed in the developing head capsule of *P. aegeria*,

where it might have become integrated with genes involved in the regulation of the optic field, such as *zen* in *Drosophila*.

The fact that some *Shx* genes appear to be deployed in different developmental contexts further complicates the use of KOs in the F<sub>0</sub>. Precise spatiotemporal control of KOs would be required to study the contributions of the *Shx* genes at different stages of development.

Disentangling the effects of the *Shx* genes at different embryological stages is crucial to study their potential pleiotropic functions. The development of transgenic techniques is therefore crucial for further functional study of the *Shx* genes in early development. Further exploiting the CRISPR/Cas9 system to produce conditional KOs is a promising avenue to produce transgenic *P. aegeria*, and is discussed in more detail below.

Understanding the role that the *Hox3* genes play in the Lepidoptera is an interesting avenue of research for several reasons. Given their fast evolution and large sequence variability both between and within species, it is likely that the different paralogs have acquired specific roles within the context of serosal specification (Ferguson *et al.*, 2014), but possibly also in the maintenance and functioning of the serosa. The duplication and fast divergence of these homeobox genes thus provides an ideal system with which to study possible neo- and sub-functionalisation events in the context of an ecologically relevant tissue (Ferguson *et al.*, 2014; Jacobs *et al.*, 2013; Jacobs *et al.*, 2014). It is possible, for example, that they allow the serosa to function in a much targeted species-specific manner depending on the ecological niche the species occupies; for example exposure to drought or high humidity, pesticides or naturally occurring toxins, parasitism and varying levels of yolk. It would mirror, to an extent, the enormous variability that can be observed in chorion (i.e. eggshell) production and morphology as a defence mechanism (e.g. chorion genes also duplicate readily and evolve very rapidly (Carter *et al.* 2013, and references therein). Furthermore, their embryonic expression requires further investigation. One could thus even ask the question whether the *Shx* genes underlie, to an extent, the ecological diversification and species-richness in the Lepidoptera. Furthermore, developing functional tools in *P. aegeria* for early developmental genes would pave the way for



large scale phylogenetic studies, such as those performed for *WntA*. The large interspecific sequence divergence and associated positive selection between species (Chapter II), suggests that the *Shx* genes have acquired species specific roles during the radiation of the Lepidoptera. Functional analyses across species are therefore necessary to elucidate the different roles the *Hox3* genes might be playing in the Lepidoptera, and will provide a better understanding of the role of transcription factor evolution in the context of duplication and divergence.

#### **IV.4.f Future Prospects**

The success in implementing CRISPR/Cas9 across such a wide range of butterfly species shows the incredible amenability of the system to study gene function. However, in order to unravel gene function in more detail, it is crucial to adapt the system to produce transgenic animals. This can be achieved by exploiting the homologous recombination mechanism (Auer *et al.*, 2014). It is possible to introduce cuts using a single or double sgRNA approach which, when coinjected with a plasmid containing the construct of interest, can insert itself endogenously into the desired locus (Ma *et al.*, 2014). This system could be exploited, for example, to produce reporter constructs that enable live visualisation of fluorescent proteins tagged to genes/enhancers of interest. Given the apparent dynamic expression of both the *Shx* genes and Wnt ligands in butterflies, this would allow for much more detailed resolution with regards to the precise spatiotemporal expression of these genes. Moreover, given that the variation responsible for wing pattern differences has been linked to clonable regions in some *Heliconius* species, the system could be used to perform enhancer swapping experiments between species, enabling the study of the contribution of enhancer evolution to wing patterning.

Furthermore, it is also theoretically possible to create overexpression constructs using the previously described *Drosophila* hsp70 promoter, which has been shown to drive expression in *B. anynana* (Tong *et al.*, 2014), a closely related species to *P. aegeria* (Peña *et al.*, 2006). This could be of particular interest to overexpress different Wnt ligands during wing development, but could also be applied to serosal development. Tagging constructs using heat

shock promoters is particularly useful since it allows for temporal control of gene editing, without the need of identifying enhancer elements, which is both technically challenging and time consuming in non-model species. This technique could allow further study of the contributions of the *Shx* genes at different stages in development.

#### **IV.4.g Conclusions**

This chapter outlines the successful implementation of CRISPR/Cas9 in my model system *P. aegeria*. Targeted mutations were introduced at four different loci, with resulting phenotypes enabling the study of their function. Mutations at the *yellow* locus showed a conserved role for this gene, resulting in reduced melanisation in affected cells. In conjunction with experiments in other butterfly species, *WntA* KO showed a conserved role for this gene in wing patterning across ~90Ma of evolution, and provides one of the first large scale functional comparative study in the Lepidoptera. Targeted mutagenesis of the *Shx* genes, however, outlines some of the limitations imposed by the technique when studying early developmental genes. It is important for future studies to address these limitations by adapting the technique for the generation of transgenic animals, allowing for the study of a wider range of developmental genes.

Chapter V

Conclusion

## V.1. *Pararge aegeria*, an emerging model species

Butterflies are considered to be very useful indicators of environmental change, particularly temperature and humidity fluctuations (Dennis, 1993). Past and current responses of butterflies to environmental variability have been well documented in terms of their numbers, life-history strategies, morphology, and biogeographical distributions (Stefanescu *et al.*, 2004). However, despite the Lepidoptera being the third largest insect order, embryological studies examining their developmental evolution are lagging behind. The available data concern themselves predominantly with the larval and adult stages, and not with what is perhaps the most vulnerable stage in the life cycle: the embryo. Most of the wealth of information recovered for the developmental evolution of Lepidoptera has historically focused on wing patterning but, with the exception of the silkworm *B. mori* (Nakao *et al.*, 2008; Nakao, 2010, 2012, 2016), little attention has been paid to their early embryonic development. Likewise, most genomic and developmental resources developed for the Lepidoptera have focused on pupal stages (e.g. (Martin and Reed, 2010b; Martin *et al.*, 2012), during which the wing discs differentiate to give rise to the variety of wing patterns observed in nature. While these data have provided spectacular examples of the evolution of Gene Regulatory Networks (GRNs) involved in regulating wing patterning (e.g. Chapter IV and Mazo-vargas *et al.*, 2017), a detailed description of early development is necessary to understand the various selective pressures that have enabled the incredibly successful radiation of the Lepidoptera.

*Pararge aegeria* has been established as a popular model system in the field of evolutionary ecology (see Chapter III and references therein). Several genomic and developmental resources have now been established for *P. aegeria*, and it is being recognised as an emerging model system in the field of evo-devo (Schmidt-Ott and Lynch, 2016). Carter *et al.* (2013) produced a detailed description of maternal effect gene expression during oogenesis resulting in a high quality annotated ovarian transcriptome, which was followed by a detailed miRNA transcriptome of the same stage (Quah *et al.*, 2015). Furthermore, it has recently been shown, using transcriptomic data, that mothers prime the development and growth of their

offspring on the basis of the environmental conditions they themselves experience during oogenesis (Gibbs and Breuker, unpublished). Moreover, a draft assembly of a genome based on next-gen illumina sequencing (Ferguson et al., 2014) and a complete mitochondrial genome (Teixeira da Costa, 2016) exist, as well as a published pipeline for transcriptome assembly and annotation (Carter *et al.*, 2016). Further to these useful genomic resources, developmental tools are also available to study the gene expression of key developmental genes through whole mount *in situ* hybridisation and antibody staining (Ferguson *et al.*, 2014; Carter, 2014). What is exciting about *P. aegeria*, is that our extensive knowledge on the evolutionary ecology of this species, including for example population and landscape differences in female reproductive strategies (e.g. Gibbs *et al.*, 2012), can now be combined with these new resources and techniques, and allow for integration into ecological evolutionary developmental studies, at a scale not possible in many classical insect model systems.

A key aspect that was still missing from the repertoire of tools available for *P. aegeria* was the development of genome editing techniques. So far, most of the data recovered for the contribution of individual genes to morphological diversity in Lepidoptera had involved detailed mapping studies and expression profiling (e.g. Belleghem *et al.*, 2017). However, few studies are available that systematically dissect the contribution of individual genes through knockdown/knockout. In part, this has been hindered by the availability of techniques that could be applied to a wide range of Lepidoptera. With this in mind, I established a protocol for targeted mutagenesis using the CRISPR/Cas9 system. As detailed in Chapter IV, I was able to introduce mutations at several loci, some of which resulted in phenotypes consistent with their hypothesised mutational effect. Knockout of the pigmentation gene *yellow* confirmed a conserved role for this gene involved in the biosynthesis of melanin (Zhang *et al.*, 2017), and allowed for effective screening of mutated individuals. Furthermore, targeting of the signalling gene *WntA*, resulted in the abolishment of several symmetry systems spanning the wing surface, confirming (along with evidence from 6 other nymphalid species) the decades worth of mapping studies implicating this gene in the evolution of wing patterning (Martin *et al.*, 2012; Martin and Reed, 2014; Jiggins *et al.*, 2017). Being able to test the functionality of specific genes, and

ultimately specific polymorphisms, makes *P. aegeria* even more attractive as not only a butterfly eco-evo-devo system, but a major insect model system in general; the knowledge and tools are now available with which to pursue research into the evolution of a wide range of developmental questions, including those relating to early embryology, even in an ecological context.

## **V.2. Duplication and Divergence at the *Hox3* Locus and the Evolutionary Success of the Lepidoptera**

*Pararge aegeria* belongs to the infraorder Heteroneura within the Ditrysia, which comprises the vast majority (98%) of Lepidopteran species (Mutanen *et al.*, 2010). Therefore, this clade largely represents both the abundant species richness of the Lepidoptera (>157,000 spp.) as well as the largest radiation of plant-feeding insects described (Bazinet *et al.*, 2013). While there are undoubtedly many factors that contributed to this enormous radiation, some key events have likely facilitated Lepidopteran evolution. Moths and butterflies most likely co-evolved in concert with the radiation of flowering plants (Mutanen *et al.*, 2010), with each species evolving specific adaptations to their host plants, in many cases requiring evolutionary innovations associated with chemical defence mechanisms (Regier *et al.*, 2015). Lepidoptera have thus evolved mechanisms to exploit and handle these chemical defences, which in many cases are plant specific (Després *et al.*, 2007; Briscoe *et al.*, 2013). However, the toxicity of host plants is one factor amongst many that contributed to the myriad of adaptations that allowed butterflies to radiate. Many species of Lepidoptera lay their eggs on the external surfaces of host plants leaving them exposed to a variety of environmental threats, including desiccation/drought resistance and pathogenic exposure. It is therefore necessary for Lepidoptera to have evolved adequate defences against these threats.

As discussed at length in Chapter I, the extraembryonic serosa appears to have evolved mechanisms to buffer insects against such environmental perturbations, through the secretion of a cuticle preventing desiccation (Rezende *et al.*, 2008; Jacobs *et al.*, 2013), and the mounting of

immune responses associated with the serosal epithelium (Jacobs *et al.*, 2014). Moreover, the serosa in collaboration with the embryo has been shown to be able to mitigate against toxins that can disrupt endogenous hormone signalling (Orth *et al.*, 2003). The evolution of these extraembryonic membranes, in turn, seems to be accompanied by changes in expression patterns and protein structure of the *Hox3/zen* genes (Chapter I Figure I – 3; Averof and Akam, 1995; Dearden *et al.*, 2000). Furthermore, it appears that the locus is amenable to tandem gene duplications (TGDs), with several insect species displaying lineage specific duplications, associated with extraembryonic tissue evolution (Stauber *et al.*, 1999; van der Zee *et al.*, 2005; Rafiqi, 2008; Ferguson *et al.*, 2014). Of these duplications, Lepidopteran *Shx* genes appear to show a vast amount of sequence divergence between both paralogous and orthologous sequences (Chapter II and Ferguson *et al.*, 2014), while retaining expression patterns consistent with serosal specification in *P. aegeria* (Carter, 2014).

Ferguson *et al.* (2014) had previously reported a conserved set of four *Shx* genes which arose during the radiation of the Ditrysia, and were able to show that periods of positive selection were associated with the paralogs following their duplication. This, as well as diverging expression patterns, indicates that the different paralogs have likely acquired new and specific functions. However, due to limited phylogenetic sampling, there were no data available for patterns of positive selection associated with specific orthologous *Shx* genes. Through estimates of amino acid substitution rates, I was able to show in Chapter II that several key residues within the divergent homeodomains of orthologous *Shx* genes have likely been under periods of positive selection, reflecting possible species specific roles. Furthermore, additional duplication events appear to have occurred during the evolution of the Ditrysia, indicating that the locus remains amenable to duplication events within the clade.

This large sequence divergence of the *Shx* genes, coupled with evidence of episodic directional selection, suggests these genes have likely evolved new roles in Ditrysiian biology. It is worthwhile noting that the serosa of Lepidoptera is somewhat unique when compared to other orders. Most other insects do not maintain this structure throughout their entire embryonic

development; it ruptures and is subsequently reabsorbed by the embryo, usually at some point following katatrepsis (Panfilio, 2008). This is in part because the extraembryonic epithelium of most insects is tightly linked to this process, where morphogenetic movements are required for embryonic rearrangements in the embryo. This does not seem to be the case in Lepidoptera, where serosal tissue remains intact almost until hatching and is not required for katatrepsis (Kobayashi, 2003). This indeed has also been observed for *P. aegeria* (Braak et al, in prep). This uncoupling of functions may have allowed Lepidoptera to maintain serosal tissue over a larger period of time, and suggests that the maintenance of this tissue is of particular significance to the order. It is possible, therefore, that TGD events in the Ditrysia facilitated asymmetric divergence of the duplicates that allowed the different homeodomains to acquire new targets involved not only in specifying, but also in the functioning and maintaining of extraembryonic membranes in a divergent way. The evolutionary pressures for the diversification in turn might have come from a need of species specific requirements in terms of desiccation resistance, toxin (including pesticides), and/or pathogenic exposure. However, it is rather remarkable that it is the transcription factors *per se* that seem to be evolving so rapidly between species. *Hox3* evolution in the Diptera also resulted in a highly divergent copy following a duplication event (*bcd*), but extensive sequence divergence between species does not exist for this particular paralog (Mcgregor, 2005). That is, when aligning *bcd* homeodomains between Dipteran species, very little nonsynonymous divergence is observed, likely due to constraints operating on downstream target recognition. In contrast, the homeodomains of the individual *Shx* genes show a large amount of variation between species (even within the relatively recently diverged *Heliconius* genus) (Chapter II).

Evidence for this kind of asymmetric divergence has been described for effector genes such as immune genes (due to the Red Queen hypothesis) (Tanaka *et al.*, 2008), but is unusual for TFs, especially homeodomain containing genes. Presumably, orthologous *Shx* genes have acquired substitutions that have enabled them to recognise different downstream targets in different Lepidopteran species. However, it is usually assumed that TFs rarely accumulate these kinds of changes, as large pleiotropic effects are expected (Carroll, 2008). A possible



explanation for this phenomenon could be due to the potential redundant functions of the *Shx* paralogs. It is possible that the *Shx* genes may be able to tolerate high levels of homeodomain substitutions if other paralogs can compensate for potential deleterious effects, effectively meaning sequence evolution is under relaxed constraint due to developmental buffering by the paralogs.

The expression patterns recovered for the *Shx* genes in *P. aegeria* do indeed argue for shared functions, as they are all expressed in overlapping patterns by the onset of zygotic transcription (Ferguson *et al.*, 2014; Carter, 2014). However, evidence for possible sub/neofunctionalisation of the paralogs also exists. Importantly, *ShxC* is the only gene to be maternally localised in the oocyte in a pattern that prefigures the location of future serosal tissue. Furthermore, *ShxD* appears to preferentially localise to a row of cells delineating the border between extraembryonic tissue and embryo proper, suggesting a key function in establishing this boundary. Likewise, *zen* expression is only observed in the follicle cells surrounding the ovarioles, and subsides until expression is detected in serosal cells following germ band retraction.

How can this large variation in homeodomain sequences be tolerated and integrated into species specific GRNs? It is interesting to note that, while there is a large amount of variation associated with the *Shx* homeodomains, the core WFQNR motif in the third alpha helix remains virtually unchanged in all the paralogs analysed. It is possible then, that target recognition does not change *per se*, but that changes outside this core motif allow for the orthologous *Shx* genes to acquire new binding affinities between species. Given that downstream targets are likely to be genes involved in immune response and or desiccation resistance, it is possible that precise fine tuning of transcription may be required in terms of dosage. Therefore, the differing affinities between orthologous *Shx* sequences may be able to mediate this fine tuning in downstream expression. Likewise, the intraspecific differences detected in *P. aegeria* within the homeodomain of some of the paralogs might be able to mediate the fine tuning of binding affinity in response to population specific pressures.

There is evidence from other TF gene families to suggest that orthologous TFs may indeed diverge in function. For example, different orthologs of yeast Fox3 exhibit substantial DNA binding diversity and are able to bind different motifs in different species (Nakagawa *et al.*, 2013). Fox3 orthologs that bind these separate motifs have been shown to have divergent substitutions in their DNA recognition helix, suggesting the binding preferences might be mediated by coding differences in orthologous Fox3 sequences. Likewise, a sea-star T-box transcription factor (Tbr), has evolved differences in motifs required for DNA binding as compared to its sea urchin ortholog, which resulted in differential recognition of binding sites between the two genes (Cheatle Jarvela and Hinman, 2015; Cheatle Jarvela and Pick, 2016).

Whether differences in orthologous *Shx* sequences have functional consequences to target recognition and/or affinity to binding sites remains to be investigated, but is a key question that could provide an insight into how TFs evolve and are incorporated into new GRNs. Firstly, it would be important to show that the paralogs are able to recognise different target sequences, through implementation of methods such as Chip-Seq (Valouev *et al.*, 2008) or in-silico modelling (cf. Ferguson *et al.*, 2014).. Using the same approach, a species comparison between orthologs could reveal potential differences in binding specificities, thus showing whether species specific binding has evolved.

### **V.3. Future Directions**

The data presented in this thesis have raised many interesting questions. Importantly, the full extent of the functionality of the *Shx* genes remains putative and is mainly based on spatio-temporal expression patterns and suggestive periods of positive selection. Conclusive evidence needs to be gathered from experiments aimed at manipulating the expression of these genes at specific times during development and possibly in specific tissues. While CRISPR/Cas9 experiments were successful for wing patterning genes, results obtained for the *Shx* genes were inconclusive and incomplete, and no direct effect of the *Shx* genes on serosal specification could be shown (Chapter IV). Ideally, follow-up experiments should take into account the potential

functional redundancy of the paralogs, and aim to knockout the *Shx* genes in various combinations, and investigate the effects across an extensive developmental time-series as they may likely be pleiotropic as well as showing an element of redundancy (see also Chapter II and IV). Furthermore, while evidence exists from other species that the serosa is directly involved in performing a variety of ecological functions ( Rezende *et al.*, 2008; Jacobs *et al.*, 2013; Jacobs *et al.*, 2014), no direct evidence is as yet available for the Lepidoptera (with the exception of a few studies in *Manduca sexta* (Berger-Twelbeck *et al.*, 2003; Orth *et al.*, 2003). It is therefore crucial for future studies to address the precise role of extraembryonic structures in the Lepidoptera, in order to make possible functional links to the diversification of the *Shx* genes.

A clue to the possible functionality of the *Shx* genes might lie in the differences in life-history traits between non- Ditrysiian and Ditrysiian Lepidoptera. Further genomic sampling at the base of the Ditrysiian split, where the evolution of the duplication of the *Shx* genes occurred, would enable further characterisation of the diversification of *Hox3/zen*. It is interesting that *Hepalius* does possess four copies of *zen*, but that these have not diversified and acquired the same characteristics of the *Shx* sequences. The common ancestor of the Hepialidae and Ditrysiia may therefore have had multiple copies of *zen*, but at which point they became diversified into new Ditrysiian specific functions is not yet known.

Overall, the evolution of the *Hox3/zen* comprises an excellent case study in TF evolution. Large phylogenetic studies have revealed that the evolution of this gene is tightly linked to the serosa, an evolutionary novelty of the pterygotes (Dearden *et al.*, 2000). Furthermore, recurrent and independent duplications have been linked to the refinement and evolution of this structure, which appears to have significant ecological roles. Whether these environmental pressures are linked to these recurrent duplications and divergence of paralogs remains to be further investigated, but the evolution and divergence of the *Shx* genes represents an ideal opportunity to answer these questions.

## Bibliography

Aalberg Haugen, I.M., Gotthard, K. (2015) Diapause induction and relaxed selection on alternative developmental pathways in a butterfly. *Journal of Animal Ecology*. **84**(2), 464–472.

Abramoff M.D., Magalhaes P.J. & Ram S.J. (2004) Image Processing with ImageJ. *Biophotonics International* **11**, 36-42.

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., Välimäki, N., Paulin, L., Kvist, J., Wahlberg, N., Tanskanen, J., Hornett, E.A., Ferguson, L.C., Luo, S., Cao, Z., Jong, M.A. de, Duplouy, A., Smolander, O.-P., Vogel, H., McCoy, R.C., Qian, K., Chong, W.S., Zhang, Q., Ahmad, F., Haukka, J.K., Joshi, A., Salojärvi, J., Wheat, C.W., Grosse-Wilde, E., Hughes, D., Katainen, R., Pitkänen, E., Ylinen, J., Waterhouse, R.M., Turunen, M., Vähärautio, A., Ojanen, S.P., Schulman, A.H., Taipale, M., Lawson, D., Ukkonen, E., Mäkinen, V., Goldsmith, M.R., Holm, L., Auvinen, P., Frilander, M.J., Hanski, I. (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications*. **5**, ncomms5737.

Alvarez W. (1972) Rotation of the Corsica–Sardinia Microplate. *Nature Physical Science* **235**,103-5.

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M., Postlethwait, J.H. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science*, **282**(5394), 1711–1714.

Arakane, Y., Dittmer, N.T., Tomoyasu, Y., Kramer, K.J., Muthukrishnan, S., Beeman, R.W., Kanost, M.R. (2010) Identification, mRNA expression and functional analysis of several *yellow* family genes in *Tribolium castaneum*. *Insect Biochemistry and Molecular Biology*. **40**(3), 259–266.

Arias, C.F., Munoz, A.G., Jiggins, C.D., Mavarez, J., Bermingham, E. and Linares, M., (2008). A hybrid zone provides evidence for incipient ecological speciation in *Heliconius* butterflies. *Molecular ecology*, *17*(21), pp.4699-4712.

Arnoult, L., Su, K.F.Y., Manoel, D., Minervino, C., Magriña, J., Gompel, N., Prud'homme, B. (2013) Emergence and Diversification of Fly Pigmentation Through Evolution of a Gene Regulatory Module. *Science*. **339**(6126), 1423–1426.

Auer, T.O., Duroure, K., Cian, A.D., Concordet, J.-P., Bene, F.D. (2014) Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair. *Genome Research*. **24**(1), 142–153.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Dharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouél, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., Wincker, P. (2006) Global

- trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*. **444**(7116), 171–178.
- Averof, M., Akam, M. (1995) Hox genes and the diversification of insect and crustacean body plans. *Nature*. **376**(6539), 420–423.
- Averof, M., Patel, N.H. (1997) Crustacean appendage evolution associated with changes in Hox gene expression. *Nature*. **388**(6643), 682–686.
- Baeg, G.H., Lin, X., Khare, N., Baumgartner, S., Perrimon, N. (2001) Heparan sulfate proteoglycans are critical for the organization of the extracellular distribution of Wingless. *Development*. **128**(1), 87–94.
- Balakirev, E.S., Anisimova, M., Ayala, F.J. (2011) Complex Interplay of Evolutionary Forces in the ladybird Homeobox Genes of *Drosophila melanogaster*. *PLOS ONE*. **6**(7), e22613.
- Balakirev, E.S., Ayala, F.J. (2004) Nucleotide Variation in the *tinman* and *bagpipe* Homeobox Genes of *Drosophila melanogaster*. *Genetics*. **166**(4), 1845–1856.
- Barton N.H. & Hewitt G.M. (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics* **16**, 113-48.
- Bassett, A.R., Liu, J.-L. (2014) CRISPR/Cas9 and Genome Editing in *Drosophila*. *Journal of Genetics and Genomics*. **41**(1), 7–19.
- Bateson, W. (1894) *Materials for the study of variation treated with especial regard to discontinuity in the origin of species*. London, New York, Macmillan and co.
- Belleghem, S.M.V., Rastas, P., Papanicolaou, A., Martin, S.H., Arias, C.F., Supple, M.A., Hanly, J.J., Mallet, J., Lewis, J.J., Hines, H.M., Ruiz, M., Salazar, C., Linares, M., Moreira, G.R.P., Jiggins, C.D., Counterman, B.A., McMillan, W.O., Papa, R. (2017) Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology & Evolution*. **1**, 0052.
- Berger-Twelbeck, P., Hofmeister, P., Emmling, S., Dorn, A. (2003) Ovicide-induced serosa degeneration and its impact on embryonic development in *Manduca sexta* (Insecta: Lepidoptera). *Tissue and Cell*. **35**(2), 101–112.
- Bergman M., Olofsson M. & Wiklund C. (2010) Contest outcome in a territorial butterfly: the role of motivation. *Proceedings of the Royal Society B: Biological Sciences*.
- Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M., Nüsslein-Volhard, C. (1988) The role of localization of bicoid RNA in organizing the anterior pattern of the *Drosophila* embryo. *The EMBO journal*. **7**(6), 1749–1756.
- Berwaerts, K., Van Dyck, H., Vints, E., Matthysen, E. (2001) Effect of manipulated wing characteristics and basking posture on thermal properties of the butterfly *Pararge aegeria* (L.). *Journal of Zoology*. **255**(2), 261–267.
- Betrán, E., Thornton, K., Long, M. (2002) Retroposed new genes out of the X in *Drosophila*. *Genome Research*. **12**(12), 1854–1859.

- Bi, H.-L., Xu, J., Tan, A.-J., Huang, Y.-P. (2016) CRISPR/Cas9-mediated targeted gene mutagenesis in *Spodoptera litura*. *Insect Science*. **23**(3), 469–477.
- Biessmann, H. (1985) Molecular analysis of the *yellow* gene (*y*) region of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*. **82**(21), 7369–7373.
- Binari, R.C., Staveley, B.E., Johnson, W.A., Godavarti, R., Sasisekharan, R., Manoukian, A.S. (1997) Genetic evidence that heparin-like glycosaminoglycans are involved in wingless signaling. *Development (Cambridge, England)*. **124**(13), 2623–2632.
- Bolker, J.A. (2008) From Embryology to Evo-Devo: A History of Developmental Evolution. *BioScience*. **58**(5), 461–463.
- Bolognesi, R., Beermann, A., Farzana, L., Wittkopp, N., Lutz, R., Balavoine, G., Brown, S.J., Schröder, R. (2008) *Tribolium* Wnts: evidence for a larger repertoire in insects with overlapping expression patterns that suggest multiple redundant functions in embryogenesis. *Development genes and evolution*. **218**(3–4), 193–202.
- Brakefield, P.M. (2007) Butterfly eyespot patterns and how evolutionary tinkering yields diversity. *Novartis Foundation Symposium*. **284**, 90-101; discussion 101-115.
- Brault, A.C., Huang, C.Y.-H., Langevin, S.A., Kinney, R.M., Bowen, R.A., Ramey, W.N., Panella, N.A., Holmes, E.C., Powers, A.M., Miller, B.R. (2007) A single positively selected West Nile viral mutation confers increased virogenesis in American crows. *Nature Genetics*. **39**(9), 1162–1166.
- Breuker C.J., Gibbs M., Merckx T., Van Dongen S. & Van Dyck H. (2010) The use of geometric morphometrics in studying butterfly wings in an evolutionary ecological context. In: *Morphometrics for non-morphometricians* (ed. by Elewa AMT), pp. DOI 10.1007/978-3-540-95853-6\_12. Springer-Verlag, Berlin Heidelberg.
- Breuker C.J., Gibbs M., Van Dyck H., Brakefield P.M., Klingenberg C.P. & Van Dongen S. (2007) Integration of wings and their eyespots in the speckled wood butterfly *Pararge aegeria*. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution* **308B**, 454-63.
- Breuker, C.J., Gibbs, M., Dongen, S.V., Merckx, T., Dyck, H.V. (2010) The Use of Geometric Morphometrics in Studying Butterfly Wings in an Evolutionary Ecological Context. In A. M. T. Elewa, ed. *Morphometrics for Nonmorphometricians*. Lecture Notes in Earth Sciences. Springer Berlin Heidelberg, pp. 271–287.
- Breuker, C.J., Gibbs, M., Van Dyck, H., Brakefield, P.M., Klingenberg, C.P., Van Dongen, S. (2007) Integration of wings and their eyespots in the speckled wood butterfly *Pararge aegeria*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. **308B**(4), 454–463.
- Briscoe, A.D. (2001) Functional Diversification of Lepidopteran Opsins Following Gene Duplication. *Molecular Biology and Evolution*. **18**(12), 2270–2279.

- Briscoe, A.D., Macias-Muñoz, A., Kozak, K.M., Walters, J.R., Yuan, F., Jamie, G.A., Martin, S.H., Dasmahapatra, K.K., Ferguson, L.C., Mallet, J., Jacquin-Joly, E., Jiggins, C.D. (2013) Female Behaviour Drives Expression and Evolution of Gustatory Receptors in Butterflies. *PLOS Genetics*. **9**(7), e1003620.
- Brunet, F.G., Roest Crollius, H., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*. **23**(9), 1808–1816.
- Bürglin, T.R., Affolter, M. (2016) Homeodomain proteins: an update. *Chromosoma*. **125**, 497–521.
- Burki, F., Kaessmann, H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nature Genetics*. **36**(10), 1061–1063.
- Calvo-Martín, J.M., Papaceit, M., Segarra, C. (2017) Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene Caf1-55 in the obscure group of *Drosophila*. *Scientific Reports*. **7**.
- Cannarozzi, G.M., Schneider, A. (2012) *Codon Evolution: Mechanisms and Models*. OUP Oxford.
- Carrasco, A.E., McGinnis, W., Gehring, W.J., Robertis, E.M.D. (1984) Cloning of an *X. laevis* gene expressed during early embryogenesis coding for a peptide region homologous to *Drosophila* homeotic genes. *Cell*. **37**(2), 409–414.
- Carroll, S.B. (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell*. **101**(6), 577–580.
- Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*. **134**(1), 25–36.
- Carroll, S.B., Gates, J., Keys, D.N., Paddock, S.W., Panganiban, G.E., Selegue, J.E., Williams, J.A. (1994) Pattern formation and eyespot determination in butterfly wings. *Science*, **265**(5168), 109–114.
- Carter J.-M. (2014) Oogenesis and Maternal Regulation of Early Development in the Speckled Wood Butterfly *Pararge aegeria*. PhD thesis, Oxford Brookes University, Oxford, U.K.
- Carter J.-M., Baker S.C., Pink R., Carter D.R.F., Collins A., Tomlin J., Gibbs M. & Breuker C.J. (2013) Unscrambling butterfly oogenesis. *BMC Genomics* **14**, 283.
- Carter, J.-M., Baker, S.C., Pink, R., Carter, D.R., Collins, A., Tomlin, J., Gibbs, M., Breuker, C.J. (2013) Unscrambling butterfly oogenesis. *BMC Genomics*. **14**, 283.
- Carter, J.-M., Gibbs, M., Breuker, C.J. (2016) Studying Oogenesis in a Non-model Organism Using Transcriptomics: Assembling, Annotating, and Analyzing Your Data. *Oogenesis*, pp.129-143.
- Carter, J.-M., Gibbs, M., Breuker, C.J. (2015) Divergent RNA Localisation Patterns of Maternal Genes Regulating Embryonic Patterning in the Butterfly *Pararge aegeria*. *PLOS ONE*. **10**(12), e0144471.

- Chai, C.-L., Zhang, Z., Huang, F.-F., Wang, X.-Y., Yu, Q.-Y., Liu, B.-B., Tian, T., Xia, Q.-Y., Lu, C., Xiang, Z.-H. (2008) A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology*. **38**(12), 1111–1120.
- Challis, R.J., Kumar, S., Dasmahapatra, K.K.K., Jiggins, C.D., Blaxter, M. (2016) Lepbase: the Lepidopteran genome database. *bioRxiv*, 056994.
- Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., Petrov, D., Jónsson, B., Schluter, D., Bell, M.A., Kingsley, D.M. (2010) Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science*. **327**(5963), 302–305.
- Chang, T., Mazotta, J., Dumstrei, K., Dumitrescu, A., Hartenstein, V. (2001) Dpp and Hh signaling in the *Drosophila* embryonic eye field. *Development (Cambridge, England)*. **128**(23), 4691–4704.
- Cheatle Jarvela, A.M., Hinman, V.F. (2015) Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *EvoDevo*. **6**, 3.
- Cheatle Jarvela, A.M., Pick, L. (2016) Evo-Devo: Discovery of Diverse Mechanisms Regulating Development. *Current Topics in Developmental Biology*. **117**, 253–274.
- Chen, B., Hrycaj, S., Schinko, J.B., Podlaha, O., Wimmer, E.A., Popadić, A., Monteiro, A. (2011) Pogostick: A New Versatile piggyBac Vector for Inducible Gene Over-Expression and Down-Regulation in Emerging Model Systems. *PLOS ONE*. **6**(4), e18659.
- Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C., Patrinos, G.P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*. **8**(10), 762–775.
- Chipman, A.D., Ferrier, D.E.K., Brena, C., Qu, J., Hughes, D.S.T., Schröder, R., Torres-Oliva, M., Znassi, N., Jiang, H., Almeida, F.C., Alonso, C.R., Apostolou, Z., Aqrabi, P., Arthur, W., Barna, J.C.J., Blankenburg, K.P., Brites, D., Capella-Gutiérrez, S., Coyle, M., Dearden, P.K., Pasquier, L.D., Duncan, E.J., Ebert, D., Eibner, C., Erikson, G., Evans, P.D., Extavour, C.G., Francisco, L., Gabaldón, T., Gillis, W.J., Goodwin-Horn, E.A., Green, J.E., Griffiths-Jones, S., Grimmelikhuijzen, C.J.P., Gubbala, S., Guigó, R., Han, Y., Hauser, F., Havlak, P., Hayden, L., Helbing, S., Holder, M., Hui, J.H.L., Hunn, J.P., Hunnekuhl, V.S., Jackson, L., Javaid, M., Jhangiani, S.N., Jiggins, F.M., Jones, T.E., Kaiser, T.S., Kalra, D., Kenny, N.J., Korchina, V., Kovar, C.L., Kraus, F.B., Lapraz, F., Lee, S.L., Lv, J., Mandapat, C., Manning, G., Mariotti, M., Mata, R., Mathew, T., Neumann, T., Newsham, I., Ngo, D.N., Ninova, M., Okwuonu, G., Onger, F., Palmer, W.J., Patil, S., Patraquim, P., Pham, C., Pu, L.-L., Putman, N.H., Rabouille, C., Ramos, O.M., Rhodes, A.C., Robertson, H.E., Robertson, H.M., Ronshaugen, M., Rozas, J., Saada, N., Sánchez-Gracia, A., Scherer, S.E., Schurko, A.M., Siggens, K.W., Simmons, D., Stief, A., Stolle, E., Telford, M.J., Tessmar-Raible, K., Thornton, R., Zee, M. van der, Haeseler, A. von, Williams, J.M., Willis, J.H., Wu, Y., Zou, X., Lawson, D., Muzny, D.M., Worley, K.C., Gibbs, R.A., Akam, M., Richards, S. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. *PLOS Biology*. **12**(11), e1002005.



- Chu, S.W., Noyes, M.B., Christensen, R.G., Pierce, B.G., Zhu, L.J., Weng, Z., Stormo, G.D., Wolfe, S.A. (2012) Exploring the DNA-recognition potential of homeodomains. *Genome Research*. **22**(10), 1889–1898.
- Clement M., Posada D. & Crandall K.A. (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657-9.
- Cohen, S.M., Brönner, G., Küttner, F., Jürgens, G., Jäckle, H. (1989) Distal-less encodes a homoeodomain protein required for limb development in *Drosophila*. *Nature*. **338**(6214), 432–434.
- Cohn, M.J., Tickle, C. (1999) Developmental basis of limblessness and axial patterning in snakes. *Nature*. **399**(6735), 474–479.
- Cooper S.J.B., Ibrahim K.M. & Hewitt G.M. (1995) Postglacial expansion and genome subdivision in the European grasshopper *Chorthippus parallelus*. *Molecular Ecology* **4**, 49-60.
- Damen, W.G., Tautz, D. (1998) A Hox class 3 orthologue from the spider *Cupiennius salei* is expressed in a Hox-gene-like fashion. *Development Genes and Evolution*. **208**(10), 586–590.
- Dapporto L. & Dennis R.L.H. (2009) Conservation biogeography of large Mediterranean islands. Butterfly impoverishment, conservation priorities and inferences for an ecological "island paradigm". *Ecography* **32**, 169-79.
- Dapporto L. (2008) Geometric morphometrics reveal male genitalia differences in the *Lasiommata megera/paramegaera* complex (Lepidoptera, Nymphalidae) and the lack of a predicted hybridization area in the Tuscan Archipelago. *Journal of Zoological Systematics and Evolutionary Research* **46**, 224-30.
- Dapporto L. (2009) Speciation in Mediterranean refugia and post-glacial expansion of *Zerynthia polyxena* (Lepidoptera, Papilionidae). *Journal of Zoological Systematics and Evolutionary Research* **48** (3), 229-237.
- Dapporto L. (2010) Satyrinae butterflies from Sardinia and Corsica show a kaleidoscopic intraspecific biogeography (Lepidoptera, Nymphalidae). *Biological Journal of the Linnean Society* **100**, 195-212.
- Dapporto L., Habel J.C., Dennis R.L.H. & Schmitt T. (2011) The biogeography of the western Mediterranean: elucidating contradictory distribution patterns of differentiation in *Maniola jurtina* (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society* **103**, 571-7.
- Dapporto, L., Bruschini, C., Dincă, V., Vila, R. and Dennis, R.L., (2012). Identifying zones of phenetic compression in West Mediterranean butterflies (Satyrinae): refugia, invasion and hybridization. *Diversity and Distributions*, **18**(11), pp.1066-1076.
- Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
- Davies N.B. (1978) Territorial defence in the speckled wood butterfly (*Pararge aegeria*): The resident always wins. *Animal Behaviour* **26, Part 1**, 138-47.

- Dearden, P., Grbic, M., Falciani, F., Akam, M. (2000) Maternal expression and early zygotic regulation of the Hox3/zen gene in the grasshopper *Schistocerca gregaria*. *Evolution & Development*. **2**(5), 261–270.
- Dehal, P., Boore, J.L. (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology*. **3**(10), e314.
- Dennis R.L.H., Shreeve T.G., Olivier A. & Coutsis J.G. (2000) Contemporary geography dominates butterfly diversity gradients within the Aegean archipelago (Lepidoptera : Papilionoidea, Hesperioidea). *Journal of Biogeography* **27**, 1365-83.
- Dennis, R.L.H. (1993) *Butterflies and Climate Change*. Manchester ; New York : New York: Manchester University Press.
- Després, L., David, J.-P., Gallet, C. (2007) The evolutionary ecology of insect resistance to plant chemicals. *Trends in Ecology & Evolution*. **22**(6), 298–307.
- deWaard J.R., Ivanova N.V., Hajibabaei M. & Hebert P.D.N. (2008) Assembling DNA Barcodes: Analytical Protocols. Pp. 275-293. In: Cristofre M. (Hrsg.), *Methods in Molecular Biology: Environmental Genetics*. Humana Press Inc., Totowa, USA, 364 pp.
- Dincă V., Wiklund C., Lukhtanov V.A., Kodandaramaiah U., Norén K., Dapporto L., Wahlberg N., Vila R. & Friberg, M. (2013) Reproductive isolation and patterns of genetic differentiation in a cryptic butterfly species complex. *Journal of Evolutionary Biology* **26**, 2095-2106
- Dincă, V., Dapporto, L. & Vila, R. (2011) A combined genetic-morphometric analysis unravels the complex biogeographic history of *Polyommatus icarus* and *P. celina* Common Blue butterflies. *Molecular Ecology* **20**, 3921-3935
- Di-Poï, N., Montoya-Burgos, J.I., Miller, H., Pourquié, O., Milinkovitch, M.C., Duboule, D. (2010) Changes in Hox genes' structure and function during the evolution of the squamate body plan. *Nature*. **464**(7285), 99–103.
- Dobson S.L., Bourtzis K., Braig H.R., Jones B.F., Zhou W.G., Rousset F. & O'Neill S.L. (1999) *Wolbachia* infections are distributed throughout insect somatic and germ line tissues. *Insect Biochemistry and Molecular Biology* **29**, 153-60.
- Dobzhansky T. (1970) *Genetics and the origin of species*. Columbia University Press, New York.
- Driever, W., Nüsslein-Volhard, C. (1988) The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell*. **54**(1), 95–104.
- Dumas P., Legeai F., Lemaitre C., Scaon E., Orsucci M., Labadie K., Gimenez S., Clamens A.L., Henri H., Vavre F., Aury J.M., Fournier P., Kergoat G.J. & d'Alençon E. (2015) *Spodoptera frugiperda* (Lepidoptera: Noctuidae) host-plant variants: two host strains or two distinct species? *Genetica* **143**, 305-16.

- Dyck, H.V., Matthysen, E. (1998) Thermoregulatory differences between phenotypes in the speckled wood butterfly: hot perchers and cold patrollers? *Oecologia*. **114**(3), 326–334.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. **32**(5), 1792–1797.
- Estrada, B., Sánchez-Herrero, E. (2001) The Hox gene Abdominal-B antagonizes appendage development in the genital disc of *Drosophila*. *Development (Cambridge, England)*. **128**(3), 331–339.
- Falciani, F., Hausdorf, B., Schröder, R., Akam, M., Tautz, D., Denell, R., Brown, S. (1996) Class 3 Hox genes in insects and the origin of zen. *Proceedings of the National Academy of Sciences*. **93**(16), 8479–8484.
- Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. 4 edition. Harlow: Pearson.
- Fawcett, J.A., Maere, S., Van de Peer, Y. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America*. **106**(14), 5737–5742.
- Ferguson L., Marletaz F., Carter J.-M., Taylor W.R., Gibbs M., Breuker C.J. & Holland P.W.H. (2014) Ancient expansion of the Hox cluster in Lepidoptera generated four Homeobox genes implicated in extraembryonic tissue formation. *PLoS Genetics* **10**, e1004698.
- Ferguson, L.C., Green, J., Surridge, A., Jiggins, C.D. (2011) Evolution of the Insect Yellow Gene Family. *Molecular Biology and Evolution*. **28**(1), 257–272.
- Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G.J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G.A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J.F., Vakhrusheva, O.A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A.S., Mark Welch, D.B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., Van Doninck, K. (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*. **500**(7463), 453–457.
- Friberg M., Vongvanich N., Borg-Karlson A.K., Kemp D.J., Merilaita S. & Wiklund C. (2008) Female mate choice determines reproductive isolation between sympatric butterflies. *Behavioral Ecology and Sociobiology* **62**, 873-86.
- Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiácovo, M.P., Church, G.M., Calarco, J.A. (2013) Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nature Methods*. **10**(8), 741–743.
- Fu, Y.X., Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*. **133**(3), 693–709.
- Fuerer, C., Habib, S.J., Nusse, R. (2010) A study on the interactions between heparan sulfate proteoglycans and Wnt proteins. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*. **239**(1), 184–190.

- Fujiwara, H., Nishikawa, H. (2016) Functional analysis of genes involved in color pattern formation in Lepidoptera. *Current Opinion in Insect Science*. **17**, 16–23.
- Futahashi, R., Kawahara-Miki, R., Kinoshita, M., Yoshitake, K., Yajima, S., Arikawa, K., Fukatsu, T. (2015) Extraordinary diversity of visual opsin genes in dragonflies. *Proceedings of the National Academy of Sciences*. **112**(11), E1247–E1256.
- Futahashi, R., Sato, J., Meng, Y., Okamoto, S., Daimon, T., Yamamoto, K., Suetsugu, Y., Narukawa, J., Takahashi, H., Banno, Y., Katsuma, S., Shimada, T., Mita, K., Fujiwara, H. (2008) *yellow* and *ebony* Are the Responsible Genes for the Larval Color Mutants of the Silkworm *Bombyx mori*. *Genetics*. **180**(4), 1995–2005.
- Gaj, T., Gersbach, C.A., Barbas III, C.F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*. **31**(7), 397–405.
- Galant, R., Carroll, S.B. (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature*. **415**(6874), 910–913.
- Gallant, J.R., Imhoff, V.E., Martin, A., Savage, W.K., Chamberlain, N.L., Pote, B.L., Peterson, C., Smith, G.E., Evans, B., Reed, R.D., Kronforst, M.R., Mullen, S.P. (2014) Ancient homology underlies adaptive mimetic diversity across butterflies. *Nature Communications*. **5**, 4817.
- Garcia-Fernández, J. (2005) The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics*. **6**(12), 881–892.
- Gaunt, S.J. (2015) The significance of Hox gene collinearity. *The International Journal of Developmental Biology*. **59**(4–6), 159–170.
- Gehring, W.J. (1985) The homeo box: A key to the understanding of development? *Cell*. **40**(1), 3–5.
- Gehring, W.J., Hiromi, Y. (1986) Homeotic genes and the homeobox. *Annual Review of Genetics*. **20**, 147–173.
- Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., Wüthrich, K. (1994) Homeodomain-DNA recognition. *Cell*. **78**(2), 211–223.
- Gibbs M. & Van Dyck H. (2009) Reproductive plasticity, oviposition site selection, and maternal effects in fragmented landscapes. *Behavioral Ecology and Sociobiology* **64**, 1-11.
- Gibbs M. & Van Dyck H. (2010) Butterfly flight activity affects reproductive performance and longevity relative to landscape structure. *Oecologia* **163**, 341-50.
- Gibbs M., Breuker C.J. & Van Dyck H. (2010b) Flight during oviposition reduces maternal egg provisioning and influences offspring development in *Pararge aegeria* (L.). *Physiological Entomology* **35**, 29-39.
- Gibbs M., Breuker C.J., Hesketh H., Hails R.S. & Van Dyck H. (2010a) Maternal effects, flight versus fecundity trade-offs, and offspring immune defence in the Speckled Wood butterfly, *Pararge aegeria*. *BMC Evolutionary Biology*, **10**.

- Gibbs M., Van Dyck H. & Breuker C.J. (2012) Development on drought-stressed host plants affects life history, flight morphology and reproductive output relative to landscape structure. *Evolutionary Applications* **5**, 66-75.
- Gibbs, A.G., Perkins, M.C., Markow, T.A. (2003) No place to hide: microclimates of Sonoran Desert *Drosophila*. *Journal of Thermal Biology*. **28**(5), 353–362.
- Gibbs, M., Breuker, C.J., Hesketh, H., Hails, R.S., Van Dyck, H. (2010) Maternal effects, flight versus fecundity trade-offs, and offspring immune defence in the Speckled Wood butterfly, *Pararge aegeria*. *BMC Evolutionary Biology*. **10**, 345.
- Gibbs, M., Breuker, C.J., Van Dyck, H. (2010) Flight during oviposition reduces maternal egg provisioning and influences offspring development in *Pararge aegeria* (L.). *Physiological Entomology*. **35**(1), 29–39.
- Gibbs, M., Dyck, H.V. (2009) Reproductive plasticity, oviposition site selection, and maternal effects in fragmented landscapes. *Behavioral Ecology and Sociobiology*. **64**(1), 1–11.
- Gibbs, M., Lace, L.A., Jones, M.J., Moore, A.J. (2004) Intraspecific competition in the speckled wood butterfly *Pararge aegeria*: Effect of rearing density and gender on larval life history. *Journal of Insect Science*. **4**(1).
- Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., Carroll, S.B. (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature*. **433**(7025), 481–487.
- Gotthard, K., Berger, D. (2010) The diapause decision as a cascade switch for adaptive developmental plasticity in body mass in a butterfly. *Journal of Evolutionary Biology*. **23**(6), 1129–1137.
- Gould, S.J. (1985) *Ontogeny and Phylogeny*. New Ed edition. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Goulson D. & Cory J.S. (1993) Flower constancy and learning in foraging preferences of the green veined butterfly *Pieris napi*. *Ecological Entomology* **18**, 315-20.
- Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J., O'Connor-Giles, K.M. (2013) Genome Engineering of *Drosophila* with the CRISPR RNA-Guided Cas9 Nuclease. *Genetics*. **194**(4), 1029–1035.
- Gratz, S.J., Rubinstein, C.D., Harrison, M.M., Wildonger, J., O'Connor-Giles, K.M. (2015) CRISPR-Cas9 genome editing in *Drosophila*. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*. **111**, 31.2.1-31.2.20.
- Grenier, J.K., Carroll, S.B. (2000) Functional evolution of the Ultrabithorax protein. *Proceedings of the National Academy of Sciences of the United States of America*. **97**(2), 704–709.
- Grens, A., Mason, E., Marsh, J.L., Bode, H.R. (1995) Evolutionary conservation of a cell fate specification gene: the Hydra achaete-scute homolog has proneural activity in *Drosophila*. *Development (Cambridge, England)*. **121**(12), 4027–4035.

- Grill A., Crnjar R., Casula P. & Menken S. (2002) Applying the IUCN threat categories to island endemics: Sardinian butterflies (Italy). *Journal for Nature Conservation* **10**, 51-60.
- Habel J.C., Dieker P. & Schmitt T. (2009) Biogeographical connections between the Maghreb and the Mediterranean peninsulas of southern Europe. *Biological Journal of the Linnean Society* **98**, 693-703.
- Habel J.C., Husemann M., Schmitt T., Dapporto L., Rodder D. & Vandewoestijne S. (2013) A Forest Butterfly in Sahara Desert Oases: Isolation Does Not Matter. *Journal of Heredity* **104**, 234-47.
- Habel J.C., Lens L., Rodder D. & Schmitt T. (2011) From Africa to Europe and back: refugia and range shifts cause high genetic differentiation in the Marbled White butterfly *Melanargia galathea*. *BMC Evolutionary Biology* **11**.
- Habel J.C., Schmitt T. & Muller P. (2005) The fourth paradigm pattern of post-glacial range expansion of European terrestrial species: the phylogeography of the Marbled White butterfly (Satyrinae, Lepidoptera). *Journal of Biogeography* **32**, 1489-97.
- Habel, J. C., Vila, R., Vodă, R., Husemann, M., Schmitt, T. and Dapporto, L. (2017), Differentiation in the marbled white butterfly species complex driven by multiple evolutionary forces. *J. Biogeogr.*, 44: 433–445.
- Halder, G., Callaerts, P., Gehring, W.J. (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science (New York, N.Y.)*. **267**(5205), 1788–1792.
- Hanes, S.D., Brent, R. (1989) DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell*. **57**(7), 1275–1283.
- Hanes, S.D., Brent, R. (1991) A genetic model for interaction of the homeodomain recognition helix with DNA. *Science (New York, N.Y.)*. **251**(4992), 426–430.
- Harrison, M.M., Jenkins, B.V., O'Connor-Giles, K.M., Wildonger, J. (2014) A CRISPR view of development. *Genes & Development*. **28**(17), 1859–1872.
- Hastings, P.J., Lupski, J.R., Rosenberg, S.M., Ira, G. (2009) Mechanisms of change in gene copy number. *Nature Reviews. Genetics*. **10**(8), 551–564.
- Heffer, A., Xiang, J., Pick, L. (2013) Variation and constraint in Hox gene evolution. *Proceedings of the National Academy of Sciences of the United States of America*. **110**(6), 2211–2216.
- Heigwer, F., Kerr, G., Boutros, M. (2014) E-CRISP: fast CRISPR target site identification. *Nature Methods*. **11**(2), 122–123.
- Hesketh, H., Gibbs, M., Breuker, C.J., Van Dyck, H., Turner, E., Hails, R.S. (2012) Exploring sub-lethal effects of exposure to a nucleopolyhedrovirus in the speckled wood (*Pararge aegeria*) butterfly. *Journal of Invertebrate Pathology*. **109**(1), 165–168.
- Hewitt G.M. (1999) Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society* **68**, 87-112.

- Hewitt G.M. (2000) The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907-13.
- Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G. & Jarvis A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-78.
- Hill J.K., Thomas C.D. & Blakeley D.S. (1999a) Evolution of flight morphology in a butterfly that has recently expanded its geographic range. *Oecologia* **121**, 165-70.
- Hill J.K., Thomas C.D. & Huntley B. (1999b) Climate and habitat availability determine 20th century changes in a butterfly's range margin. *Proceedings of the Royal Society of London Series B-Biological Sciences* **266**, 1197-206.
- Hoekstra, H.E., Coyne, J.A. (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution; International Journal of Organic Evolution*. **61**(5), 995–1016.
- Hof, A.E. van't, Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., Saccheri, I.J. (2016) The industrial melanism mutation in British peppered moths is a transposable element. *Nature*. **534**(7605), 102–105.
- Holland, P.W., Garcia-Fernández, J., Williams, N.A., Sidow, A. (1994) Gene duplications and the origins of vertebrate development. *Development (Cambridge, England). Supplement*, 125–133.
- Holland, P.W., Hogan, B.L. (1988) Expression of homeo box genes during mouse development: a review. *Genes & Development*. **2**(7), 773–782.
- Holland, P.W.H. (2013) Evolution of homeobox genes. *Wiley Interdisciplinary Reviews. Developmental Biology*. **2**(1), 31–45.
- Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L., Paps, J. (2017) New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. **372**(1713).
- Horn, T., Hilbrant, M., Panfilio, K.A. (2015) Evolution of epithelial morphogenesis: phenotypic integration across multiple levels of biological organization. *Frontiers in Genetics*. **6**.
- Hrycaj, S.M., Wellik, D.M. (2016) Hox genes and evolution. *F1000Research*. **5**.
- Hsu, P.D., Lander, E.S., Zhang, F. (2014) Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*. **157**(6), 1262–1278.
- Hufnagel, L., Kreuger, J., Cohen, S.M., Shraiman, B.I. (2006) On the role of glypicans in the process of morphogen gradient formation. *Developmental Biology*. **300**(2), 512–522.
- Hughes, C.L., Kaufman, T.C. (2002) Hox genes and the evolution of the arthropod body plan. *Evolution & Development*. **4**(6), 459–499.
- Hughes, C.L., Liu, P.Z., Kaufman, T.C. (2004) Expression patterns of the rogue Hox genes Hox3/zen and fushi tarazu in the apterygote insect *Thermobia domestica*. *Evolution & Development*. **6**(6), 393–401.

- Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., Joung, J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology*. **31**(3), 227–229.
- Innan, H., Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*. **11**(2), 97–108.
- Jacob, F. (1977) Evolution and tinkering. *Science*. **196**(4295), 1161–1166.
- Jacob, F. (1993) *The Logic of Life: A History of Heredity*. New e. edition. Princeton, N.J.: Princeton University Press.
- Jacobs, C.G.C., Rezende, G.L., Lamers, G.E.M., Zee, M. van der (2013) The extraembryonic serosa protects the insect egg against desiccation. *Proceedings of the Royal Society of London B: Biological Sciences*. **280**(1764), 20131082.
- Jacobs, C.G.C., Spaink, H.P., Zee, M. van der (2014) The extraembryonic serosa is a frontier epithelium providing the insect egg with a full-range innate immune response. *eLife*. **3**, e04111.
- Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., de Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J., Roest Crollius, H. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*. **431**(7011), 946–957.
- Jiggins, C.D., Wallbank, R.W.R., Hanly, J.J. (2017) Waiting in the wings: what can we learn about gene co-option from the diversification of butterfly wing patterns? *Phil. Trans. R. Soc. B*. **372**(1713), 20150485.
- Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A. (2014) Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science*. **343**(6176), 1247997.
- Joga, M.R., Zotti, M.J., Smaghe, G., Christiaens, O. (2016) RNAi Efficiency, Systemic Properties, and Novel Delivery Methods for Pest Insect Control: What We Know So Far. *Frontiers in Physiology*. **7**.
- Johnson, H., Solensky, M.J., Satterfield, D.A., Davis, A.K. (2014) Does Skipping a Meal Matter to a Butterfly's Appearance? Effects of Larval Food Stress on Wing Morphology and Color in Monarch Butterflies. *PLOS ONE*. **9**(4), e93492.
- Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*. **131**(3), 530–543.



- Jovelin, R., Dunham, J.P., Sung, F.S., Phillips, P.C. (2009) High Nucleotide Divergence in Developmental Regulatory Genes Contrasts With the Structural Elements of Olfactory Pathways in *Caenorhabditis*. *Genetics*. **181**(4), 1387–1397.
- Kemp, D.J., Wiklund, C., Dyck, H.V. (2006) Contest behaviour in the speckled wood butterfly (*Pararge aegeria*): seasonal phenotypic plasticity and the functional significance of flight performance. *Behavioral Ecology and Sociobiology*. **59**(3), 403–411.
- Kenny, N.J., Chan, K.W., Nong, W., Qu, Z., Maeso, I., Yip, H.Y., Chan, T.F., Kwan, H.S., Holland, P.W.H., Chu, K.H., Hui, J.H.L. (2016) Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity*. **116**(2), 190–199.
- Ketmaier V., Giusti F. & Caccone A. (2006) Molecular phylogeny and historical biogeography of the land snail genus *Solatopupa* (Pulmonata) in the peri-Tyrrhenian area. *Molecular Phylogenetics and Evolution* **39**, 439-51.
- Keys, D.N., Lewis, D.L., Selegue, J.E., Pearson, B.J., Goodrich, L.V., Johnson, R.L., Gates, J., Scott, M.P., Carroll, S.B. (1999) Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science (New York, N.Y.)*. **283**(5401), 532–534.
- Klingenberg, C.P. (1998) Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews of the Cambridge Philosophical Society*. **73**(1), 79–123.
- Klingenberg, C.P. (2011) MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*. **11**(2), 353–357.
- Klingenberg, C.P., McIntyre, G.S. (1998) Geometric Morphometrics of Developmental Instability: Analyzing Patterns of Fluctuating Asymmetry with Procrustes Methods. *Evolution*. **52**(5), 1363–1375.
- Kolliopoulou, A., Swevers, L. (2014) Recent progress in RNAi research in Lepidoptera: intracellular machinery, antiviral immune response and prospects for insect pest control. *Current Opinion in Insect Science*. **6**, 28–34.
- Kondrashov, F.A. (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings. Biological Sciences*. **279**(1749), 5048–5057.
- Koutroumpa, F.A., Monsempes, C., François, M.-C., de Cian, A., Royer, C., Concordet, J.-P., Jacquin-Joly, E. (2016) Heritable genome editing with CRISPR/Cas9 induces anosmia in a crop pest moth. *Scientific Reports*. **6**.
- Kozak, K.M., Wahlberg, N., Neild, A.F.E., Dasmahapatra, K.K., Mallet, J., Jiggins, C.D. (2015) Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic *Heliconius* Butterflies. *Systematic Biology*. **64**(3), 505–524.
- Kronforst, M.R., Papa, R. (2015) The functional basis of wing patterning in *Heliconius* butterflies: the molecules behind mimicry. *Genetics*. **200**(1), 1–19.
- Kumar S., Stecher G. & Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870-4.

- Kumar, S., Stecher, G., Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*. **33**(7), 1870–1874.
- Kuznetsova A., Bruun Brockhoff P. & R. H.B.C. (2016) lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). r package version 2.0-33.
- Lamer, A., Dorn, A. (2001) The serosa of *Manduca sexta* (Insecta, Lepidoptera): ontogeny, secretory activity, structural changes, and functional considerations. *Tissue and Cell*. **33**(6), 580–595.
- Leigh J.W. & Bryant D. (2015) POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**, 1110-6.
- Levine, M., Rubin, G.M., Tjian, R. (1984) Human DNA sequences homologous to a protein coding region conserved between homeotic genes of *Drosophila*. *Cell*. **38**(3), 667–673.
- Levinson, G., Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*. **4**(3), 203–221.
- Lewis, E.B. (1978) A gene complex controlling segmentation in *Drosophila*. *Nature*. **276**(5688), 565–570.
- Lewis, E.B. (1994) Homeosis: the first 100 years. In *Genes, Development, and Cancer*. Springer, Dordrecht, pp. 505–510.
- Lewis, E.B. (1998) The bithorax complex: the first fifty years. *The International Journal of Developmental Biology*. **42**(3), 403–415.
- Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., Barker, M.S. (2015) Early genome duplications in conifers and other seed plants. *Science Advances*. **1**(10), e1501084.
- Librado P. & Rozas J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2.
- Liu, J., Lemonds, T.R., Marden, J.H., Popadić, A. (2016) A Pathway Analysis of Melanin Patterning in a Hemimetabolous Insect. *Genetics*. **203**(1), 403–413.
- Löhr, U., Pick, L. (2005) Cofactor-interaction motifs and the cooption of a homeotic Hox protein into the segmentation pathway of *Drosophila melanogaster*. *Current biology: CB*. **15**(7), 643–649.
- Löhr, U., Yussa, M., Pick, L. (2001) *Drosophila fushi tarazu*. a gene on the border of homeotic function. *Current biology: CB*. **11**(18), 1403–1412.
- Long, M., Langley, C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science (New York, N.Y.)*. **260**(5104), 91–95.
- Longdon B., Day J., Schulz N., Leftwich P.T., de Jong M.A., Breuker C.J., Gibbs M., Obbard D.J., Wilfert L., Smith S.C., McGonigle J.E., Houslay T.M., Wright L.I., Livraghi L., Evans L.C., Friend L.A., Chapman T., Vontas J., Kambouraki N. & Jiggins F.M. (2017) Vertically

transmitted rhabdoviruses are found across three insect families and have dynamic interactions with their hosts. *Proceedings of the Royal Society B-Biological Sciences* **284**, 20162381.

Longdon B., Murray G.G., Palmer W.J., Day J.P., Parker D.J., Welch J.J., Obbard D.J. & Jiggins F.M. (2015) The evolution, diversity, and host associations of rhabdoviruses. *Virus Evol* **1**, vev014.

Lynch, M., Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*. **290**(5494), 1151–1155.

Ma, L.-J., Ibrahim, A.S., Skory, C., Grabherr, M.G., Burger, G., Butler, M., Elias, M., Idnurm, A., Lang, B.F., Sone, T., Abe, A., Calvo, S.E., Corrochano, L.M., Engels, R., Fu, J., Hansberg, W., Kim, J.-M., Kodira, C.D., Koehrsen, M.J., Liu, B., Miranda-Saavedra, D., O’Leary, S., Ortiz-Castellanos, L., Poulter, R., Rodriguez-Romero, J., Ruiz-Herrera, J., Shen, Y.-Q., Zeng, Q., Galagan, J., Birren, B.W., Cuomo, C.A., Wickes, B.L. (2009) Genomic Analysis of the Basal Lineage Fungus *Rhizopus oryzae* Reveals a Whole-Genome Duplication. *PLOS Genetics*. **5**(7), e1000549.

Ma, S., Chang, J., Wang, X., Liu, Y., Zhang, J., Lu, W., Gao, J., Shi, R., Zhao, P., Xia, Q. (2014) CRISPR/Cas9 mediated multiplex genome editing and heritable mutagenesis of BmKu70 in *Bombyx mori*. *Scientific Reports*. **4**, 4489.

Macqueen, D.J., Johnston, I.A. (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. R. Soc. B*. **281**(1778), 20132881.

Makova, K.D., Li, W.-H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Research*. **13**(7), 1638–1645.

Mann, R.S. (1997) Why are Hox genes clustered? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*. **19**(8), 661–664.

Mann, R.S., Chan, S.K. (1996) Extra specificity from extradenticle: the partnership between HOX and PBX/EXD homeodomain proteins. *Trends in genetics: TIG*. **12**(7), 258–262.

Mann, R.S., Lelli, K.M., Joshi, R. (2009) Hox specificity unique roles for cofactors and collaborators. *Current Topics in Developmental Biology*. **88**, 63–101.

Marcus, J.M., Ramos, D.M., Monteiro, A. (2004) Germline transformation of the butterfly *Bicyclus anynana*. *Proceedings. Biological Sciences*. **271 Suppl 5**, S263-265.

Markert, M.J., Zhang, Y., Enuameh, M.S., Reppert, S.M., Wolfe, S.A., Merlin, C. (2016) Genomic Access to Monarch Migration Using TALEN and CRISPR/Cas9-Mediated Targeted Mutagenesis. *G3: Genes, Genomes, Genetics*. **6**(4), 905–915.

Martens, C., Van de Peer, Y. (2010) The hidden duplication past of the plant pathogen *Phytophthora* and its consequences for infection. *BMC Genomics*. **11**, 353.

Martin, A., Papa, R., Nadeau, N.J., Hill, R.I., Counterman, B.A., Halder, G., Jiggins, C.D., Kronforst, M.R., Long, A.D., McMillan, W.O., Reed, R.D. (2012) Diversification of complex butterfly wing patterns by repeated regulatory evolution of a Wnt ligand. *Proceedings of the National Academy of Sciences of the United States of America*. **109**(31), 12632–12637.

- Martin, A., Reed, R.D. (2010a) wingless and aristaless2 Define a Developmental Ground Plan for Moth and Butterfly Wing Pattern Evolution. *Molecular Biology and Evolution*. **27**(12), 2864–2878.
- Martin, A., Reed, R.D. (2014) Wnt signaling underlies evolution and development of the butterfly wing pattern symmetry systems. *Developmental Biology*. **395**(2), 367–378.
- Mauffret A., Contrucci I. & Brunet C. (1999) Structural evolution of the Northern Tyrrhenian Sea from new seismic data. *Marine and Petroleum Geology* **16**, 381-407.
- Mazo-Vargas A., Concha C., Livraghi L., Massardo D., Wallbank R., Zhang L., Papador J., Martinez-Najera D., Jiggins C., Kronforst M., Breuker C.J., Reed R., Patel N., McMillan W., Martin A. (2017). Macro-evolutionary shifts of WntA function potentiate butterfly wing pattern diversity. *Proceedings of the National Academy of Sciences*, 114(40):10701–10706.
- McDonald, J.H., Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*. **351**(6328), 652–654.
- McGinnis, W., Garber, R.L., Wirz, J., Kuroiwa, A., Gehring, W.J. (1984) A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell*. **37**(2), 403–408.
- McGinnis, W., Krumlauf, R. (1992) Homeobox genes and axial patterning. *Cell*. **68**(2), 283–302.
- McMillan, W.O., Monteiro, A., Kapan, D.D. (2002) Development and evolution on the wing. *Trends in Ecology & Evolution*. **17**(3), 125–133.
- Meiklejohn, C.D., Coolon, J.D., Hartl, D.L., Wittkopp, P.J. (2014) The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Research*. **24**(1), 84–95.
- Mendel, G., Punnett, R.C., Burndy Library, donor D. (1866) *Versuche über Pflanzen-Hybriden*. Brünn : Im Verlage des Vereines.
- Merckx T. & Van Dyck H. (2006) Landscape structure and phenotypic plasticity in flight morphology in the butterfly *Pararge aegeria*. *Oikos* **113**, 226-32.
- Merrill, R.M., Dasmahapatra, K.K., Davey, J.W., Dell’Aglia, D.D., Hanly, J.J., Huber, B., Jiggins, C.D., Joron, M., Kozak, K.M., Llaurens, V., Martin, S.H., Montgomery, S.H., Morris, J., Nadeau, N.J., Pinharanda, A.L., Rosser, N., Thompson, M.J., Vanjari, S., Wallbank, R.W.R., Yu, Q. (2015) The diversification of *Heliconius* butterflies: what have we learned in 150 years? *Journal of Evolutionary Biology*. **28**(8), 1417–1438.
- Meulenkamp J.E. & Sissingh W. (2003) Tertiary palaeogeography and tectonostratigraphic evolution of the Northern and Southern Peri-Tethys platforms and the intermediate domains of the African-Eurasian convergent plate boundary zone. *Palaeogeography Palaeoclimatology Palaeoecology* **196**, 209-28.
- Meyer, A., Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*. **27**(9), 937–945.

- Mittmann, B., Wolff, C. (2012) Embryonic development and staging of the cobweb spider *Parasteatoda tepidariorum* C. L. Koch, 1841 (syn.: *Achaearanea tepidariorum*; Araneomorphae; Theridiidae). *Development Genes and Evolution*. **222**(4), 189–216.
- Monod, J.Y. (1988) *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. New York: Random House USA Inc.
- Monteiro, A., Chen, B., Ramos, D.M., Oliver, J.C., Tong, X., Guo, M., Wang, W.-K., Fazzino, L., Kamal, F. (2013) Distal-less regulates eyespot patterns and melanization in *Bicyclus* butterflies. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*. **320**(5), 321–331.
- Müller, M.M., Carrasco, A.E., DeRobertis, E.M. (1984) A homeo-box-containing gene expressed during oogenesis in xenopus. *Cell*. **39**(1), 157–162.
- Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., Scheffler, K. (2013) FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular Biology and Evolution*. **30**(5), 1196–1205.
- Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., Pond, S.L.K. (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLOS Genetics*. **8**(7), e1002764.
- Mutanen, M., Wahlberg, N., Kaila, L. (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20100392.
- Nadeau, N.J., Pardo-Diaz, C., Whibley, A., Supple, M.A., Saenko, S.V., Wallbank, R.W.R., Wu, G.C., Maroja, L., Ferguson, L., Hanly, J.J., Hines, H., Salazar, C., Merrill, R.M., Dowling, A.J., French-Constant, R.H., Llaurens, V., Joron, M., McMillan, W.O., Jiggins, C.D. (2016) The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*. **534**(7605), 106–110.
- Nakagawa, S., Gisselbrecht, S.S., Rogers, J.M., Hartl, D.L., Bulyk, M.L. (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences*. **110**(30), 12349–12354.
- Nakao, H. (2010) Characterization of *Bombyx* embryo segmentation process: expression profiles of engrailed, even-skipped, caudal, and wnt1/wingless homologues. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. **314B**(3), 224–231.
- Nakao, H. (2012) Anterior and posterior centers jointly regulate *Bombyx* embryo body segmentation. *Developmental Biology*. **371**(2), 293–301.
- Nakao, H. (2016) Hunchback knockdown induces supernumerary segment formation in *Bombyx*. *Developmental Biology*. **413**(2), 207–216.
- Nakao, H., Matsumoto, T., Oba, Y., Niimi, T., Yaginuma, T. (2008) Germ cell specification and early embryonic patterning in *Bombyx mori* as revealed by nanos orthologues. *Evolution & Development*. **10**(5), 546–554.

- Nielsen, R. ed. (2005) *Statistical Methods in Molecular Evolution*. 2005 edition. New York: Springer.
- Nijhout, H.F. (1978) Wing pattern formation in Lepidoptera: A model. *Journal of Experimental Zoology*. **206**(2), 119–136.
- Nijhout, H.F., Maini, P.K., Madzvamuse, A., Wathen, A.J., Sekimura, T. (2003) Pigmentation pattern formation in butterflies: experiments and models. *Comptes Rendus Biologies*. **326**(8), 717–727.
- Nüsslein-Volhard, C., Wieschaus, E. (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature*. **287**(5785), 795–801.
- Nylin, S., Wickman, P.-O., Wiklund, C. (1995) Life-cycle regulation and life history plasticity in the speckled wood butterfly: are reaction norms predictable? *Biological Journal of the Linnean Society*. **55**(2), 143–157.
- Ohno, S. (2013) *Evolution by Gene Duplication*. Springer Science & Business Media.
- Okonechnikov, K., Golosova, O., Fursov, M. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. **28**(8), 1166–1167.
- Oliver T.H., Marshall H.H., Morecroft M.D., Brereton T., Prudhomme C. & Huntingford C. (2015) Interacting effects of climate change and habitat fragmentation on drought-sensitive butterflies. *Nature Climate Change* **5**, 941-5.
- Orth, A.P., Tauchman, S.J., Doll, S.C., Goodman, W.G. (2003) Embryonic expression of juvenile hormone binding protein and its relationship to the toxic effects of juvenile hormone in *Manduca sexta*. *Insect Biochemistry and Molecular Biology*. **33**(12), 1275–1284.
- Otaki, J.M. (2011) Color-Pattern Analysis of Eyespots in Butterfly Wings: A Critical Examination of Morphogen Gradient Models. *Zoological Science*. **28**(6), 403–413.
- Panfilio, K.A. (2008) Extraembryonic development in insects and the acrobatics of blastokinesis. *Developmental Biology*. **313**(2), 471–491.
- Panfilio, K.A., Akam, M. (2007) A comparison of Hox3 and Zen protein coding sequences in taxa that span the Hox3/zen divergence. *Development Genes and Evolution*. **217**(4), 323–329.
- Papa, R., Kapan, D.D., Counterman, B.A., Maldonado, K., Lindstrom, D.P., Reed, R.D., Nijhout, H.F., Hrbek, T., McMillan, W.O. (2013) Multi-Allelic Major Effect Genes Interact with Minor Effect QTLs to Control Adaptive Color Pattern Variation in *Heliconius erato*. *PLOS ONE*. **8**(3), e57033.
- Papillon, D., Telford, M.J. (2007) Evolution of Hox3 and ftz in arthropods: insights from the crustacean *Daphnia pulex*. *Development Genes and Evolution*. **217**(4), 315–322.
- Parchem, R.J., Perry, M.W., Patel, N.H. (2007) Patterns on the insect wing. *Current Opinion in Genetics & Development*. **17**(4), 300–308.
- Parmesan C. (1999) Metapopulation ecology. *Nature* **399**, 747.

- Pascual-Anaya, J., D'Aniello, S., Kuratani, S., Garcia-Fernández, J. (2013) Evolution of Hox gene clusters in deuterostomes. *BMC developmental biology*. **13**, 26.
- Patrick Callaerts, Georg Halder, Gehring, and W.J. (1997) Pax-6 in Development and Evolution. *Annual Review of Neuroscience*. **20**(1), 483–532.
- Pellegroms, B., Van Dongen, S., Van Dyck, H., Lens, L. (2009) Larval food stress differentially affects flight morphology in male and female speckled woods (*Pararge aegeria*). *Ecological Entomology*. **34**(3), 387–393.
- Peña, C., Wahlberg, N., Weingartner, E., Kodandaramaiah, U., Nylin, S., Freitas, A.V.L., Brower, A.V.Z. (2006) Higher level phylogeny of Satyrinae butterflies (Lepidoptera: Nymphalidae) based on DNA sequence data. *Molecular Phylogenetics and Evolution*. **40**(1), 29–49.
- Perry, M., Kinoshita, M., Saldi, G., Huo, L., Arikawa, K., Desplan, C. (2016) Molecular logic behind the three-way stochastic choices that expand butterfly colour vision. *Nature*. **535**(7611), 280–284.
- Pick, L. (2016) Hox genes, evo-devo, and the case of the ftz gene. *Chromosoma*. **125**(3), 535–551.
- Pinzari M. & Sbordoni V. (2013) Species and mate recognition in two sympatric Grayling butterflies: *Hipparchia fagi* and *H. hermione genava* (Lepidoptera). *Ethology Ecology & Evolution* **25**, 28-51.
- Platt, R.J., Chen, S., Zhou, Y., Yim, M.J., Swiech, L., Kempton, H.R., Dahlman, J.E., Parnas, O., Eisenhaure, T.M., Jovanovic, M., Graham, D.B., Jhunjhunwala, S., Heidenreich, M., Xavier, R.J., Langer, R., Anderson, D.G., Hacohen, N., Regev, A., Feng, G., Sharp, P.A., Zhang, F. (2014) CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling. *Cell*. **159**(2), 440–455.
- Pond, K., L, S., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W., Scheffler, K. (2011) A Random Effects Branch-Site Model for Detecting Episodic Diversifying Selection. *Molecular Biology and Evolution*. **28**(11), 3033–3043.
- Pond, Kosakovsky, L, S., Frost, S.D.W. (2005) Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution*. **22**(5), 1208–1222.
- Pond, S.L.K., Frost, S.D.W. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*. **21**(10), 2531–2533.
- Pond, Sergei L. Kosakovsky, Frost, S.D.W., Muse, S.V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. **21**(5), 676–679.
- Presgraves D.C., Balagopalan L., Abmayr S.M. & Orr H.A. (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* **423**, 715-9.
- Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.-D., True, J.R., Carroll, S.B. (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*. **440**(7087), 1050–1053.

QGIS Development Team (2009) QGIS Geographic Information System. Open Source Geospatial Foundation. URL <http://qgis.osgeo.org>.

Qi, L., Fang, Q., Zhao, L., Xia, H., Zhou, Y., Xiao, J., Li, K., Ye, G. (2016) De Novo Assembly and Developmental Transcriptome Analysis of the Small White Butterfly *Pieris rapae*. *PLOS ONE*. **11**(7), e0159258.

Quah S., Breuker C.J. & Holland P.W. (2015) A Diversity of Conserved and Novel Ovarian MicroRNAs in the Speckled Wood (*Pararge aegeria*). *PloS one* **10**, e0142243.

Quah, S., Breuker, C.J., Holland, P.W.H. (2015) A Diversity of Conserved and Novel Ovarian MicroRNAs in the Speckled Wood (*Pararge aegeria*). *PLOS ONE*. **10**(11), e0142243.

Quiring, R., Walldorf, U., Kloter, U., Gehring, W.J. (1994) Homology of the eyeless gene of *Drosophila* to the Small eye gene in mice and Aniridia in humans. *Science (New York, N.Y.)*. **265**(5173), 785–789.

R Development Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rafiqi, A.M. (2008) *Morphological transitions and the genetic basis of the evolution of extraembryonic tissues in flies*. s.n.].

Ramos, D.M., Kamal, F., Wimmer, E.A., Cartwright, A.N., Monteiro, A. (2006) Temporal and spatial control of transgene expression using laser induction of the hsp70 promoter. *BMC developmental biology*. **6**, 55.

Ramos, D.M., Monteiro, A. (2007) Transgenic approaches to study wing color pattern development in Lepidoptera. *Molecular bioSystems*. **3**(8), 530–535.

Reams, A.B., Roth, J.R. (2015) Mechanisms of Gene Duplication and Amplification. *Cold Spring Harbor Perspectives in Biology*. **7**(2).

Rebeiz, M., Williams, T.M. (2017) Using *Drosophila* pigmentation traits to study the mechanisms of cis-regulatory evolution. *Current Opinion in Insect Science*. **19**, 1–7.

Reed, R.D., Papa, R., Martin, A., Hines, H.M., Counterman, B.A., Pardo-Diaz, C., Jiggins, C.D., Chamberlain, N.L., Kronforst, M.R., Chen, R., Halder, G., Nijhout, H.F., McMillan, W.O. (2011) optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science (New York, N.Y.)*. **333**(6046), 1137–1141.

Regier, J.C., Mitter, C., Kristensen, N.P., Davis, D.R., Van Nieuwerkerken, E.J., Rota, J., Simonsen, T.J., Mitter, K.T., Kawahara, A.Y., Yen, S.-H., Cummings, M.P., Zwick, A. (2015) A molecular phylogeny for the oldest (nonditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. *Systematic Entomology*. **40**(4), 671–704.

Rezende, G.L., Martins, A.J., Gentile, C., Farnesi, L.C., Pelajo-Machado, M., Peixoto, A.A., Valle, D. (2008) Embryonic desiccation resistance in *Aedes aegypti*: presumptive role of the chitinized Serosal Cuticle. *BMC Developmental Biology*. **8**, 82.



- Ribera I. & Volger A.P. (2004) Speciation of Iberian diving beetles in Pleistocene refugia (Coleoptera, Dytiscidae). *Molecular Ecology* **13** (1), 179-193.
- Richter-Boix A., Quintela M., Kierczak M., Franch M. & Laurila A. (2013) Fine-grained adaptive divergence in an amphibian: genetic basis of phenotypic divergence and the role of nonrandom gene flow in restricting effective migration among wetlands. *Molecular Ecology* **22**, 1322-40.
- Rogers S.M. & Bernatchez L. (2006) The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*). *Journal of Evolutionary Biology* **19**, 1979-94.
- Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*. **461**(7268), 1248–1253.
- Ronshaugen, M., McGinnis, N., McGinnis, W. (2002) Hox protein mutation and macroevolution of the insect body plan. *Nature*. **415**(6874), 914–917.
- Rozas, J. (2009) DNA sequence polymorphism analysis using DnaSP. *Methods in Molecular Biology (Clifton, N.J.)*. **537**, 337–350.
- Rushlow, C., Levine, M. (1990) Role Of The *zerknüllt* Gene In Dorsal-Ventral Pattern Formation In *Drosophila*. *Advances in Genetics*. **27**, 277–307.
- Sawyer, S.L., Wu, L.I., Emerman, M., Malik, H.S. (2005) Positive selection of primate TRIM5 $\alpha$  identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*. **102**(8), 2832–2837.
- Schmidt-Ott, U. (2000) The amnioserosa is an apomorphic character of cyclorrhaphan flies. *Development Genes and Evolution*. **210**(7), 373–376.
- Schmidt-Ott, U., Lynch J.A. (2016) Emerging developmental genetic model systems in holometabolous insects. *Current Opinion in Genetics & Development*. **39**, 116-128.
- Schmidt-Ott, U., Kwan, C.W. (2016) Morphogenetic functions of extraembryonic membranes in insects. *Current Opinion in Insect Science*. **13**, 86–92.
- Schmidt-Ott, U., Rafiqi, A.M., Lemke, S. (2010) Hox3/zen and the Evolution of Extraembryonic Epithelia in Insects. In J. S. Deutsch, ed. *Hox Genes*. Advances in Experimental Medicine and Biology. Springer New York, pp. 133–144.
- Schmitt T. (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology* **4**.
- Schneider, C.A., Rasband, W.S., Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*. **9**(7), 671–675.
- Schwager, E.E., Sharma, P.P., Clarke, T., Leite, D.J., Wierschin, T., Pechmann, M., Akiyama-Oda, Y., Esposito, L., Bechsgaard, J., Bilde, T., Buffry, A.D., Chao, H., Dinh, H., Doddapaneni, H., Dugan, S., Eibner, C., Extavour, C.G., Funch, P., Garb, J., Gonzalez, V.L., Griffiths-Jones, S., Han, Y., Hayashi, C., Hilbrant, M., Hughes, D.S.T., Janssen, R., Lee, S.L., Maeso, I.,

Murali, S.C., Muzny, D.M., Fonseca, R.N. da, Qu, J., Ronshaugen, M., Schomburg, C., Schoenauer, A., Stollewerk, A., Torres-Oliva, M., Turetzek, N., Vanthournout, B., Werren, J., Wolff, C., Worley, K.C., Bucher, G., Gibbs, R.A., Coddington, J., Oda, H., Stanke, M., Ayoub, N.A., Prpic, N.-M., Flot, J.-F., Posnien, N., Richards, S., McGregor, A.P. (2017) The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *bioRxiv*, 106385.

Schwanwitsch, B.N. (1924) 21. On the Ground-plan of Wing-pattern in Nymphalids and certain other Families of the Rhopalocerous Lepidoptera. *Proceedings of the Zoological Society of London*. **94**(2), 509–528.

Seddon J.M., Santucci F., Reeve N.J. & Hewitt G.M. (2001) DNA footprints of European hedgehogs, *Erinaceus europaeus* and *E. concolor*. Pleistocene refugia, postglacial expansion and colonization routes. *Molecular Ecology* **10**, 2187-98.

Shubin, N., Tabin, C., Carroll, S. (2009) Deep homology and the origins of evolutionary novelty. *Nature*. **457**(7231), 818–823.

Shukla, J.N., Kalsi, M., Sethi, A., Narva, K.E., Fishilevich, E., Singh, S., Mogilicherla, K., Palli, S.R. (2016) Reduced stability and intracellular transport of dsRNA contribute to poor RNAi response in lepidopteran insects. *RNA Biology*. **13**(7), 656–669.

Sibly, R.M., Winokur, L., Smith, R.H. (1997) Interpopulation Variation in Phenotypic Plasticity in the Speckled Wood Butterfly, *Pararge aegeria*. *Oikos*. **78**(2), 323–330.

Skelhorn, J., Holmes, G.G., Hossie, T.J., Sherratt, T.N. (2016) Eyespots. *Current biology: CB*. **26**(2), R52-54.

Soares, A.E.R., Soares, M.A., Schrago, C.G. (2008) Positive Selection on HIV Accessory Proteins and the Analysis of Molecular Adaptation After Interspecies Transmission. *Journal of Molecular Evolution*. **66**(6), 598.

Spaethe, J., Briscoe, A.D. (2004) Early Duplication and Functional Diversification of the Opsin Gene Family in Insects. *Molecular Biology and Evolution*. **21**(8), 1583–1594.

Stauber, M., Jäckle, H., Schmidt-Ott, U. (1999) The anterior determinant bicoid of *Drosophila* is a derived Hox class 3 gene. *Proceedings of the National Academy of Sciences*. **96**(7), 3786–3789.

Stauber, M., Prell, A., Schmidt-Ott, U. (2002) A single Hox3 gene with composite bicoid and zerknüllt expression characteristics in non-Cyclorrhaphan flies. *Proceedings of the National Academy of Sciences*. **99**(1), 274–279.

Stefanescu, C., Herrando, S., Páramo, F. (2004) Butterfly species richness in the north-west Mediterranean Basin: the role of natural and human-induced factors. *Journal of Biogeography*. **31**(6), 905–915.

Stern, D.L. (2000) Evolutionary developmental biology and the problem of variation. *Evolution; International Journal of Organic Evolution*. **54**(4), 1079–1091.

- Stern, D.L., Frankel, N. (2013) The structure and evolution of *cis*-regulatory regions: the *shavenbaby* story. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. **368**(1632), 20130028.
- Stern, D.L., Orgogozo, V. (2009) Is genetic evolution predictable? *Science (New York, N.Y.)*. **323**(5915), 746–751.
- Taberlet P., Fumagalli L., Wust-Saucy A.G. & Cosson J.F. (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453–64.
- Tajima, F. (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*. **123**(3), 585–595.
- Takasu, Y., Kobayashi, I., Beumer, K., Uchino, K., Sezutsu, H., Sajwan, S., Carroll, D., Tamura, T., Zurovec, M. (2010) Targeted mutagenesis in the silkworm *Bombyx mori* using zinc finger nuclease mRNA injection. *Insect Biochemistry and Molecular Biology*. **40**(10), 759–765.
- Takasu, Y., Sajwan, S., Daimon, T., Osanai-Futahashi, M., Uchino, K., Sezutsu, H., Tamura, T., Zurovec, M. (2013) Efficient TALEN Construction for *Bombyx mori* Gene Targeting. *PLOS ONE*. **8**(9), e73458.
- Takasu, Y., Tamura, T., Goldsmith, M., Zurovec, M. (2016) Targeted Mutagenesis in *Bombyx mori* Using TALENs. In R. Kühn, W. Würst, & B. Wefers, eds. *TALENs*. Methods in Molecular Biology. Springer New York, pp. 127–142.
- Talloon, W., Dyck, H.V., Lens, L. (2004) The Cost of Melanization: Butterfly Wing Coloration Underenvironmental Stress. *Evolution*. **58**(2), 360–366.
- Tamura, T., Thibert, C., Royer, C., Kanda, T., Eappen, A., Kamba, M., Kômoto, N., Thomas, J.-L., Mauchamp, B., Chavancy, G., Shirk, P., Fraser, M., Prudhomme, J.-C., Couble, P. (2000) Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. *Nature Biotechnology*. **18**(1), 81–84.
- Tanaka, H., Ishibashi, J., Fujita, K., Nakajima, Y., Sagisaka, A., Tomimoto, K., Suzuki, N., Yoshiyama, M., Kaneko, Y., Iwasaki, T., Sunagawa, T., Yamaji, K., Asaoka, A., Mita, K., Yamakawa, M. (2008) A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochemistry and Molecular Biology*. **38**(12), 1087–1110.
- Teixeira da Costa, L.F. (2016) The complete mitochondrial genome of *Parage aegeria* (Insecta: Lepidoptera: Papilionidae). *Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis*. **27**(1), 551–552.
- Templeton A.R., Routman E. & Phillips C.A. (1995) Separating population structure from population history: a cladistic analysis of geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–82.
- Terenius, O., Papanicolaou, A., Garbutt, J.S., Eleftherianos, I., Huvenne, H., Kanginakudru, S., Albrechtsen, M., An, C., Aymeric, J.-L., Barthel, A., Bebas, P., Bitra, K., Bravo, A., Chevalier, F., Collinge, D.P., Crava, C.M., de Maagd, R.A., Duvic, B., Erlandson, M., Faye, I., Felföldi, G., Fujiwara, H., Futahashi, R., Gandhe, A.S., Gatehouse, H.S., Gatehouse, L.N., Giebertowicz, J.M., Gómez, I., Grimmelikhuijzen, C.J.P., Groot, A.T., Hauser, F., Heckel, D.G., Hegedus,

D.D., Hrycaj, S., Huang, L., Hull, J.J., Iatrou, K., Iga, M., Kanost, M.R., Kotwica, J., Li, C., Li, J., Liu, J., Lundmark, M., Matsumoto, S., Meyering-Vos, M., Millichap, P.J., Monteiro, A., Mrinal, N., Niimi, T., Nowara, D., Ohnishi, A., Oostra, V., Ozaki, K., Papakonstantinou, M., Popadic, A., Rajam, M.V., Saenko, S., Simpson, R.M., Soberón, M., Strand, M.R., Tomita, S., Toprak, U., Wang, P., Wee, C.W., Whyard, S., Zhang, W., Nagaraju, J., French-Constant, R.H., Herrero, S., Gordon, K., Swevers, L., Smagghe, G. (2011) RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. *Journal of Insect Physiology*. **57**(2), 231–245.

Ting, C.-T., Tsaur, S.-C., Sun, S., Browne, W.E., Chen, Y.-C., Patel, N.H., Wu, C.-I. (2004) Gene duplication and speciation in *Drosophila*: Evidence from the Odysseus locus. *Proceedings of the National Academy of Sciences of the United States of America*. **101**(33), 12232–12235.

Ting, C.-T., Tsaur, S.-C., Wu, M.-L., Wu, C.-I. (1998) A Rapidly Evolving Homeobox at the Site of a Hybrid Sterility Gene. *Science*. **282**(5393), 1501–1504.

Tison J.-L., Edmark V.N., Sandoval-Castellanos E., Van Dyck H., Tammaru T., Välimäki P., Dalén L. & Gotthard K. (2014) Signature of post-glacial expansion and genetic structure at the northern range limit of the speckled wood butterfly. *Biological Journal of the Linnean Society* **113**, 136-48.

Toews D.P.L. & Brelsford A. (2012) The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology* **21**, 3907-30.

Tong, X., Hrycaj, S., Podlaha, O., Popadic, A., Monteiro, A. (2014) Over-expression of Ultrabithorax alters embryonic body plan and wing patterns in the butterfly *Bicyclus anynana*. *Developmental Biology*. **394**(2), 357–366.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., Desplan, C. (1989) A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell*. **59**(3), 553–562.

Vachon, G., Cohen, B., Pfeifle, C., McGuffin, M.E., Botas, J., Cohen, S.M. (1992) Homeotic genes of the Bithorax complex repress limb development in the abdomen of the *Drosophila* embryo through the target gene Distal-less. *Cell*. **71**(3), 437–450.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*. **5**(9), 829–834.

Van Dongen, S. and Talloen, W. (2007) Phenotypic and genetic variations and correlations in multitrait developmental instability: a multivariate Bayesian model applied to Speckled Wood butterfly (*Pararge aegeria*) wing measurements. *Genetics Research*, **89**(3), 155-163.

Van der Geer A., Lyras G., de Vos J. & Dermitzakis M. (2010). Evolution of island mammals: adaptation and extinction of placental mammals on islands. 479 p. Wiley Blackwell, UK.

van der Zee, M., Berns, N., Roth, S. (2005) Distinct Functions of the *Tribolium zerknüllt* Genes in Serosa Specification and Dorsal Closure. *Current Biology*. **15**(7), 624–636.

Van Dyck, H., Matthysen, E., Dhondt, A.A. (1997) The effect of wing colour on male behavioural strategies in the speckled wood butterfly. *Animal Behaviour*. **53**(1), 39–51.

- Van Dyck, H., Wiklund, C. (2002) Seasonal butterfly design: morphological plasticity among three developmental pathways relative to sex, flight and thermoregulation. *Journal of Evolutionary Biology*. **15**(2), 216–225.
- van Schooten, B., Jiggins, C.D., Briscoe, A.D., Papa, R. (2016) Genome-wide analysis of ionotropic receptors provides insight into their evolution in *Heliconius* butterflies. *BMC genomics*. **17**, 254.
- Vitti, J.J., Grossman, S.R., Sabeti, P.C. (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*. **47**, 97–120.
- Vodă R., Dapporto L. Dincă V. & Vila, R. (2015a) Cryptic matters: overlooked species generate most butterfly beta-diversity. *Ecography* **38**(4), 405-409.
- Voda R., Dapporto L., Dincă V. & Vila R. (2015b) Why do cryptic species tend not to co-occur? A case study on two cryptic pairs of butterflies. *PLoS ONE* **10**(2), e0117802.
- Vodă, R., Dapporto, L., Dincă, V., Shreeve, T.G., Khaldi, M., Barech, G., Rebbas, K., Sammut, P., Scalercio, S., Hebert, P.D. and Vila, R., 2016. Historical and contemporary factors generate unique butterfly communities on islands. *Scientific Reports*, **6**.
- Waaaijers, S., Portegijs, V., Kerver, J., Lemmens, B.B.C.G., Tijsterman, M., Heuvel, S. van den, Boxem, M. (2013) CRISPR/Cas9-Targeted Mutagenesis in *Caenorhabditis elegans*. *Genetics*, genetics.113.156299.
- Wahlberg, N., Leneuve, J., Kodandaramaiah, U., Peña, C., Nylin, S., Freitas, A.V.L., Brower, A.V.Z. (2009) Nymphalid butterflies diversify following near demise at the cretaceous/tertiary boundary. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20091303.
- Wakimoto, B.T., Turner, F.R., Kaufman, T.C. (1984) Defects in embryogenesis in mutants associated with the antennapedia gene complex of *Drosophila melanogaster*. *Developmental Biology*. **102**(1), 147–172.
- Wallbank, R.W.R., Baxter, S.W., Pardo-Diaz, C., Hanly, J.J., Martin, S.H., Mallet, J., Dasmahapatra, K.K., Salazar, C., Joron, M., Nadeau, N., McMillan, W.O., Jiggins, C.D. (2016) Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling. *PLoS Biology*. **14**(1), e1002353.
- Wang, Y., Li, Z., Xu, J., Zeng, B., Ling, L., You, L., Chen, Y., Huang, Y., Tan, A. (2013) The CRISPR/Cas system mediates efficient genome engineering in *Bombyx mori*. *Cell Research*. **23**(12), 1414–1416.
- Wanner, K.W., Robertson, H.M. (2008) The gustatory receptor family in the silkworm moth *Bombyx mori* is characterized by a large expansion of a single lineage of putative bitter receptors. *Insect Molecular Biology*. **17**(6), 621–629.
- Warren, R.W., Nagy, L., Selegue, J., Gates, J., Carroll, S. (1994) Evolution of homeotic gene regulation and function in flies and butterflies. *Nature*. **372**(6505), 458–461.
- Waters, J.M., 2011. Competitive exclusion: phylogeography’s „elephant in the room“?. *Molecular Ecology*, **20**(21), pp.4388-4394.

- Weatherbee, S.D., Nijhout, H.F., Grunert, L.W., Halder, G., Galant, R., Selegue, J., Carroll, S. (1999) Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. *Current Biology*. **9**(3), 109–115.
- Weingartner E., Wahlberg N. & Nylin S. (2006) Speciation in *Pararge* (Satyrinae : Nymphalidae) butterflies - North Africa is the source of ancestral populations of all *Pararge* species. *Systematic Entomology* **31**, 621-32.
- Werner, T., Koshikawa, S., Williams, T.M., Carroll, S.B. (2010) Generation of a novel wing colour pattern by the Wingless morphogen. *Nature*. **464**(7292), 1143–1148.
- Wilfert L. & Jiggins F.M. (2014) Flies on the move: an inherited virus mirrors *Drosophila melanogaster*'s elusive ecology and demography. *Molecular Ecology* **23**, 2093-104.
- Wittkopp, P.J., Beldade, P. (2009) Development and evolution of insect pigmentation: Genetic mechanisms and the potential consequences of pleiotropy. *Seminars in Cell & Developmental Biology*. **20**(1), 65–71.
- Wittkopp, P.J., True, J.R., Carroll, S.B. (2002) Reciprocal functions of the *Drosophila* yellow and ebony proteins in the development and evolution of pigment patterns. *Development (Cambridge, England)*. **129**(8), 1849–1858.
- Wittkopp, P.J., Vaccaro, K., Carroll, S.B. (2002) Evolution of yellow Gene Regulation and Pigmentation in *Drosophila*. *Current Biology*. **12**(18), 1547–1556.
- Wolfe, K.H., Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. **387**(6634), 708–713.
- Woltering, J.M., Vonk, F.J., Müller, H., Bardine, N., Tuduice, I.L., de Bakker, M.A.G., Knöchel, W., Sirbu, I.O., Durston, A.J., Richardson, M.K. (2009) Axial patterning in snakes and caecilians: Evidence for an alternative interpretation of the Hox code. *Developmental Biology*. **332**(1), 82–89.
- Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., Xu, J., Zheng, X., Ren, L., Wang, G., Zhang, Y., Huo, L., Zhao, Z., Cao, D., Lu, C., Li, C., Zhou, Y., Liu, Z., Fan, Z., Shan, G., Li, X., Wu, S., Song, Lipu, Hou, G., Jiang, Y., Jeney, Z., Yu, D., Wang, L., Shao, C., Song, Lai, Sun, J., Ji, P., Wang, Jian, Li, Q., Xu, L., Sun, F., Feng, J., Wang, C., Wang, S., Wang, B., Li, Y., Zhu, Y., Xue, W., Zhao, L., Wang, Jintu, Gu, Y., Lv, W., Wu, K., Xiao, J., Wu, J., Zhang, Z., Yu, J., Sun, X. (2014) Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genetics*. **46**(11), 1212–1219.
- Yang, Z., Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*. **15**(12), 496–503.
- Zeh, D.W., Zeh, J.A., Smith, R.L. (1989) Ovipositors, Amnions and Eggshell Architecture in the Diversification of Terrestrial Arthropods. *The Quarterly Review of Biology*. **64**(2), 147–168.
- Zhang, L., Martin, A., Perry, M.W., Burg, K.R.L. van der, Matsuoka, Y., Monteiro, A., Reed, R.D. (2017) Genetic Basis of Melanin Pigmentation in Butterfly Wings. *Genetics*. **205**(4), 1537–1550.

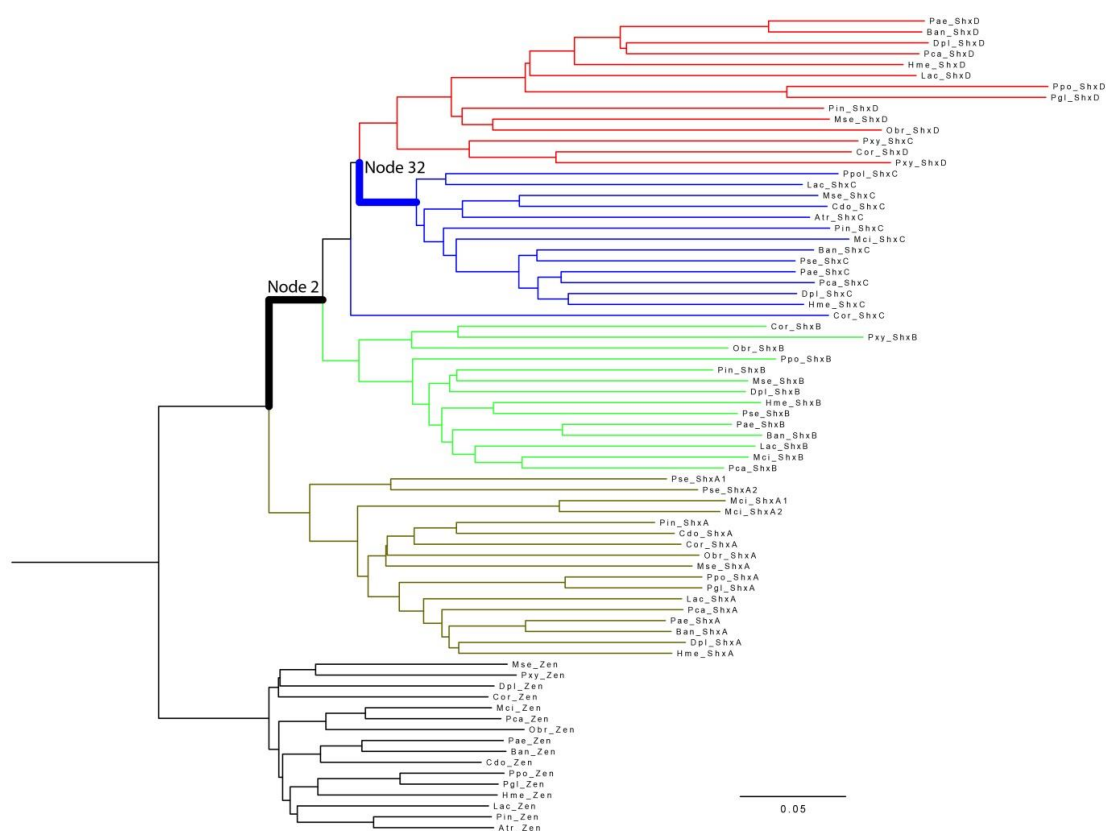
Zhang, L., Reed, R.D. (2016) Genome editing in butterflies reveals that spalt promotes and Distal-less represses eyespot colour patterns. *Nature Communications*. **7**, 11769.

Zinetti F., Dapporto L., Vovlas A., Chelazzi G., Bonelli S., Balletto E., Ciofi C. (2013) When the rule becomes the exception. No evidence of gene flow between two *Zerynthia* cryptic butterflies suggests the emergence of a new model group. *PLoS ONE* **8**(6), e65746.

**Appendices**



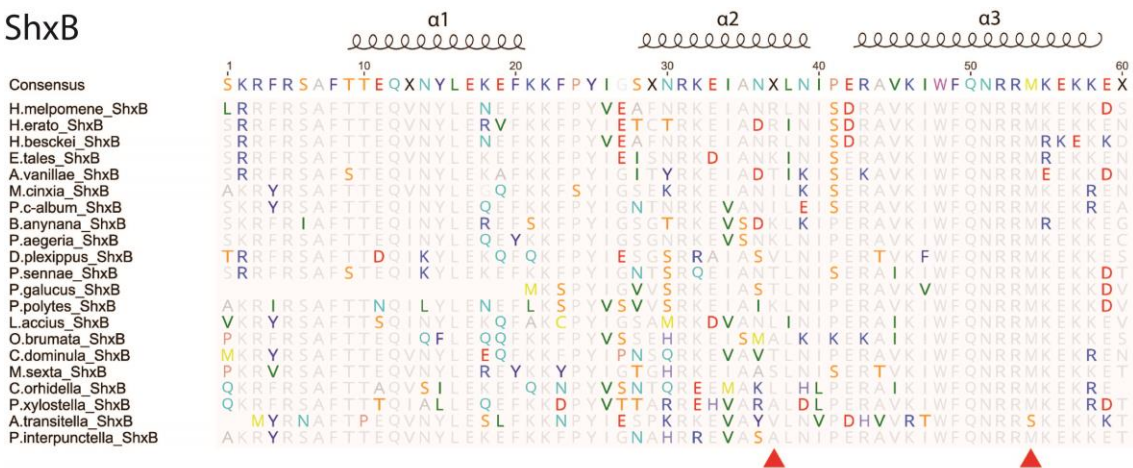
## Appendix II



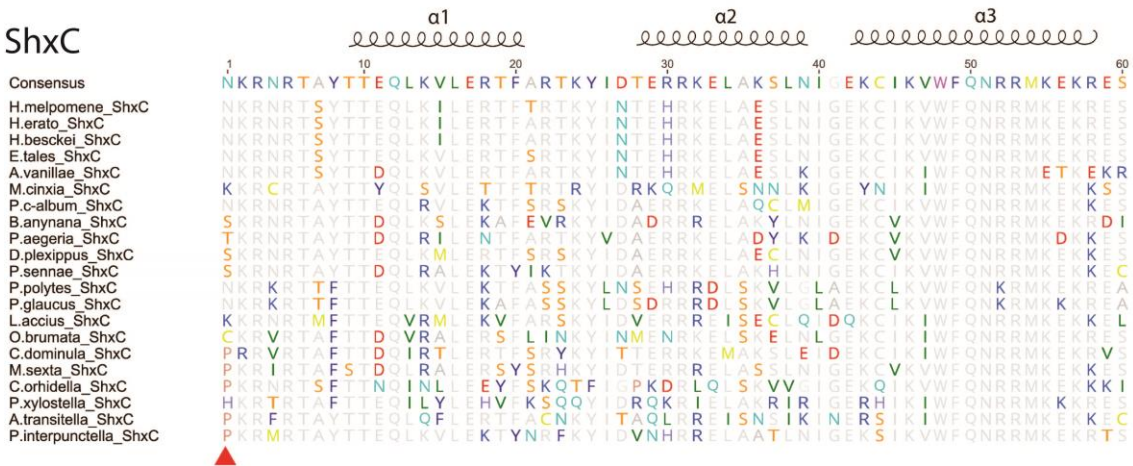
**Appendix II - Figure 1. Tree topology based on Homeodomain alignments used in selection analyses (BS-REL).**

Thick lines correspond to branches inferred to be under positive selection (see Table II- 1). A best fitting nucleotide substitution model was obtained by using the maximum likelihood model approach in MEGA7. A Maximum likelihood tree was obtained from a homeodomain alignment using RAxML and LG+ $\Gamma$  model with 100 bootstrap replicates for analyses spanning all Lepidoptera. The topology inferred from these trees was used in subsequent BS-REL.

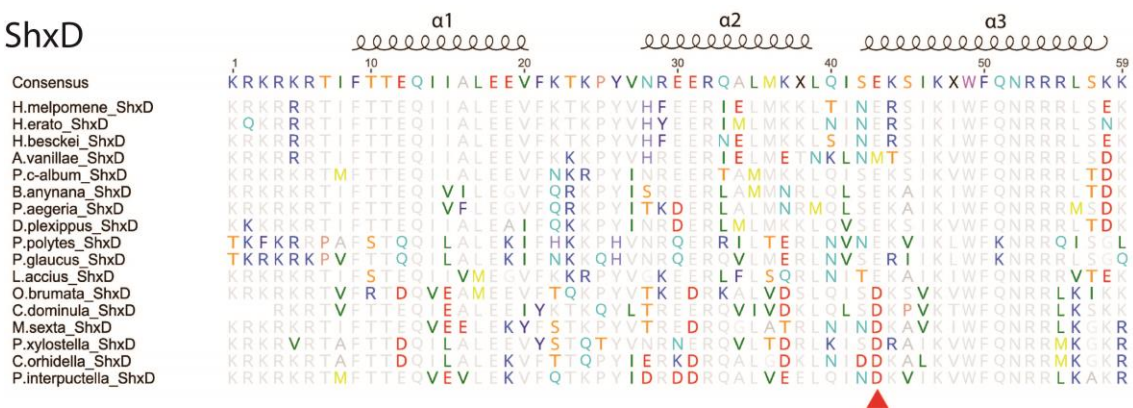
## ShxB



## ShxC



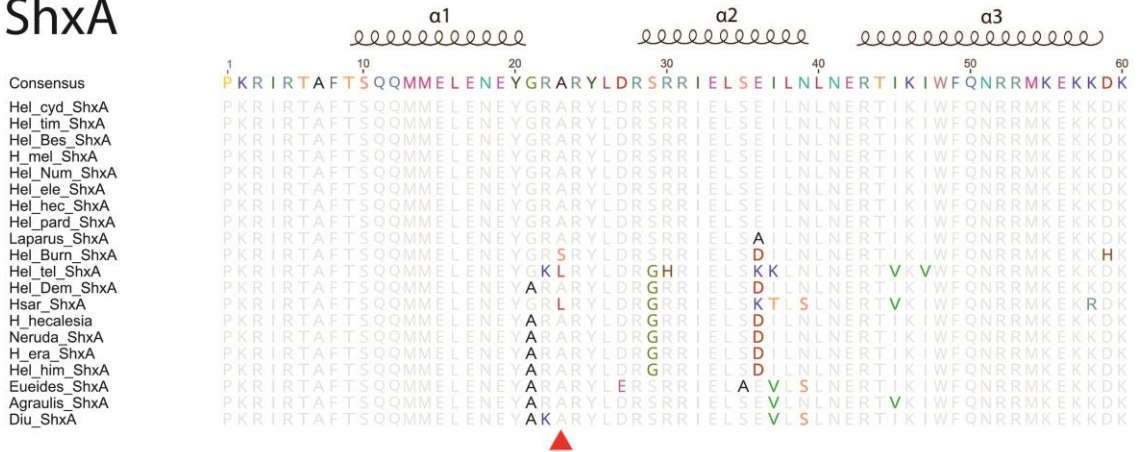
## ShxD



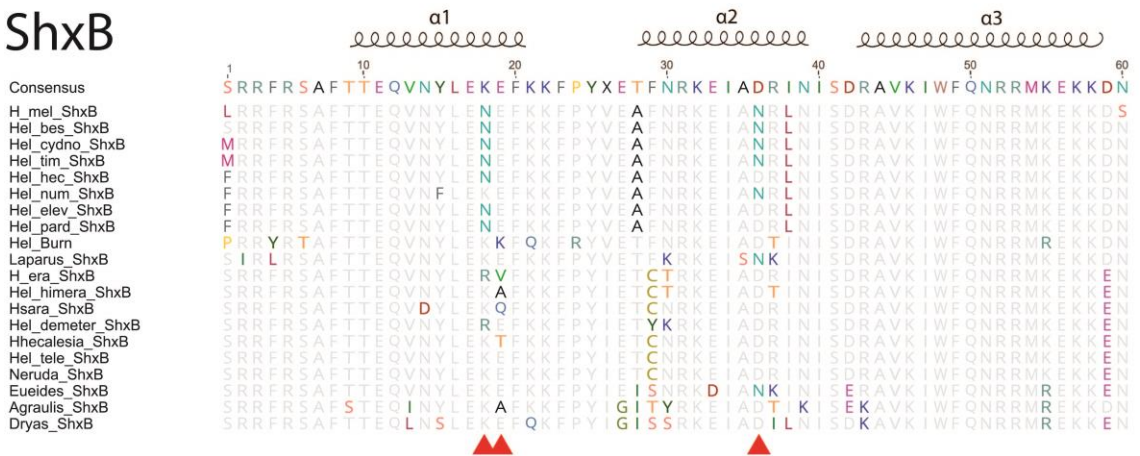
### Appendix II – Figure 2. Putative location of positively selected sites as identified by MEME on homeodomain alignments spanning all Lepidoptera.

Location of positively selected sites is indicated by red arrows. Alpha helix position is inferred by homology to insect Antp protein, for which a crystal structure is available.

## ShxA



## ShxB



## ShxD



### Appendix II – Figure 3. Putative location of positively selected sites as identified by MEME on homeodomain alignments spanning *Heliconius*.

Location of positively selected sites is indicated by red arrows. Alpha helix position is inferred by homology to insect Antp protein, for which a crystal structure is available.

**Appendix II - Table 1. Homeodomain replacement polymorphisms in *Pararge aegeria***

Homeodomain position and respective replacement polymorphism are shown. Replacement from the most common amino acid at each position and its respective codon change is outlined.

Gene	HD Position	Freq	Codon Change	Amino Acid Replacement
<i>ShxA</i>	39	1	AAC -> AGC	Asn -> Ser
<i>ShxB</i>	19	6	TAT -> TTT	Tyr -> Phe
	44	1	GTA -> TTA	Val -> Leu
<i>ShxD</i>	4	11	AAG -> ACG	Lys -> Thr
	15	3	CTC -> TTC	Leu -> Phe
	47	2	ATT -> GTT	Ile -> Val
	55	36	ATC -> ATG	Ile -> Met

**Appendix II - Table 2. Homeodomain replacement polymorphisms in *Heliconius erato***

Homeodomain position and respective replacement polymorphism are shown. Replacement from the most common amino acid at each position and its respective codon change is outlined.

Gene	HD Position	Freq	Codon Change	Amino Acid Replacement
<i>ShxA</i>	34	4	TTA -> ATA	Leu -> Ile
<i>ShxB</i>	1	47	TGG -> TCG	Ser -> Trp
		44	TGG -> ACG	Ser -> Thr
		9	TGG -> AGG	Ser -> Arg
	17	23	GCT -> TCG	Ala -> Ser
		13	GCT -> TAT	Ala -> Tyr
		12	GCT -> GAG	Ala -> Glu
		5	GCT -> GAT	Ala -> Asp
	18	58	AAG -> AGG	Lys -> Arg
	19	40	GCC -> GTA	Ala -> Val
		5	GCC -> AAA	Ala -> Lys
		2	GCC -> ATC	Ala -> Ile
	21	3	AAA -> CAA	Lys -> Gln
	22	6	AAG -> AGG	Lys -> Arg
		3	AAG -> ATG	Lys -> Met
	23	4	TTT -> TCT	Phe -> Ser
	33	2	GAA -> CAA	Gln -> Glu
	36	45	GAT -> GAA	Asp -> Glu
	37	24	AGA -> ACA	Arg -> Thr
		6	AGA -> AAA	Arg -> Lys
42	4	GAC -> GAG	Asp -> Glu	
59	3	GAT -> GAA	Asp -> Glu	
60	2	AAT -> AGT	Asn -> Ser	
<i>ShxC</i>	1	3	AAC -> AAG	Asn -> Lys
	21	3	GCC -> ACC	Ala -> Thr
	32	2	AAA -> GAA	Lys -> Glu
	39	46	AAA -> AAT	Lys -> Asn
	42	2	GAA -> CAA	Glu -> Gln
	43	6	AAA -> AAT	Lys -> Asn
	47	2	GTA -> TTA	Val -> Leu
	54	2	ATG -> GTG	Met -> Val
<i>ShxD</i>	17	3	GAA -> AAA	Glu -> Lys
	27	19	CAT -> GAT	His -> Asp
	28	2	TAT -> TTT	Tyr -> Phe
	33	42	CGG -> CTG	Arg -> Leu
		16	CGG -> ATG	Arg -> Met
	2	CGG -> AAG	Arg -> Lys	

		2	CGG -> CAG	Arg -> Gln
35		56	ATT -> ATG	Ile -> Met
37		7	AAA -> AGA	Lys -> Arg
57		69	AAT -> GAT	Asn -> Asp
		11	AAT -> GCT	Asn -> Ala
		4	AAT -> TAT	Asn -> Tyr
		2	AAT -> TTT	Asn -> Ile
60		15	AGT -> AAT	Ser -> Asn
<hr/>				
<i>Zen</i>	22	28	CAA -> GAA	Gln -> Glu
	39	59	CAA -> CAC	Gln -> His
<hr/>				

**Appendix II - Table 3. Annotation of the *Hox3* locus in the Leiodopteran species analysed**

Presence/absence of paralogs is shown, as well as partial annotation where complete CDS could not be recovered. Partial sequences were not used in the analyses.

Species	<i>ShxA</i>	<i>ShxB</i>	<i>ShxC</i>	<i>ShxD</i>	<i>Zen</i>
<i>M. cinxia</i>	Annotated (two copies)	Annotated	Annotated	Absent (?)	Annotated
<i>B. anynana</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>D. plexippus</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>P. sennae</i>	Annotated (two copies)	Annotated	Annotated	Absent	Annotated
<i>P. galucus</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>P. polytes</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>P. xuthus</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>L. accius</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>O. brumata</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>M. sexta</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>A. transitella</i>	Annotated (17 copies?)	Annotated (two copies)	Annotated	Region of N's (absent?)	Annotated
<i>P. interpunctella</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>A. vanillae</i>	Annotated	Annotated	Annotated	Partial	Annotated
<i>E. tales</i>	Annotated	Annotated	Annotated	Partial	Annotated
<i>H. besckei</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. burneyi</i>	Annotated	Annotated	Partial	Annotated	Annotated
<i>H. cydno</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. demeter</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. elevatus</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. erato</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. hecale</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. himera</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. numata</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. pardalinus</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. telesiphe</i>	Annotated	Annotated	Annotated	Partial	Annotated
<i>H. timareta</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>L. doris</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>N. aode</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. hecalesia</i>	Annotated	Annotated	Annotated	Annotated	Annotated
<i>H. sara</i>	Annotated	Annotated	Annotated	Partial	Annotated
<i>D. iulia</i>	Annotated	Annotated	Annotated	Annotated	Annotated

**Appendix II- Table 4.**

Summary of MEME analysis on Lepidopteran *Shx* genes.

Homeodomains. Average  $\omega$  is shown next to gene.

<b>MEME Homeodomain Across Leps</b>					
<b>ShxA:</b> Average $\omega$ = 0.065 (95% CI [0.055,0.077])					
<b>ShxB:</b> Average $\omega$ = 0.13 (95% CI [0.11,0.14])					
Codon	$\alpha$	$\beta$ -	Pr[ $\beta$ = $\beta$ -]	p-value	q-value
54	0	0	0.931646	0.050037	1
37	0.443229	0	0.499098	0.043213	1
<b>ShxC:</b> Average $\omega$ = 0.12 (95% CI [0.10,0.14])					
Codon	$\alpha$	$\beta$ -	Pr[ $\beta$ = $\beta$ -]	p-value	q-value
1	0.306002	0.171591	0.846797	0.053247	1
<b>ShxD:</b> Average $\omega$ = 0.13 (95% CI [0.12,0.15])					
Codon	$\alpha$	$\beta$ -	Pr[ $\beta$ = $\beta$ -]	p-value	q-value
43	0.150693	0	0.859757	0.047098	1
<b>Zen:</b> Average $\omega$ = 0.0038 (95% CI [0.0015,0.0077])					

<b>MEME SUMMARY</b>						
Gene	HD Position	$\alpha$	$\beta$ -	Pr[ $\beta$ = $\beta$ -]	p-value	
ShxB	54	0	0	0.93	0.05	
	37	0.44	0	0.50	0.04	
ShxC	1	0.31	0.17	0.85	0.05	
ShxD	43	0.15	0	0.86	0.05	

**Appendix II – Table 5.**

Summary of sites under selection in *Heliconius* using all methods. Significant sites at  $p=0.1$  are indicated in bold. Asterisks denote homeodomain positions.

HELICONIUS RADIATION DATA						
ShxA						
Codon	SLAC	FEL	IFEL	REL	MEME	FUBAR
12	0.640 (0.199)	<b>1.093 (0.077)</b>	0(1)	-0.150 (10.135)	> <b>100 (0.100)</b>	0.107 (0.739)
20	0.321 (0.444)	0.461 (0.145)	<b>1.36 (0.07)</b>	-0.123 (9.593)	> <b>100 (0.067)</b>	0.067 (0.681)
34	0.366 (0.453)	1.304 (0.234)	0(1)	-0.311 (4.184)	> <b>100 (0.077)</b>	0.046 (0.598)
38	0.004 (0.683)	0.369 (0.688)	<b>-0.83 (0.35)</b>	-0.364 (2.822)	<b>10.203 (0.076)</b>	0.003 (0.473)
49	0.625 (0.303)	<b>1.333 (0.052)</b>	<b>3.17 (0.02)</b>	-0.101 (13.992)	> <b>100 (0.100)</b>	0.143 (0.789)
55	0.307 (0.479)	0.607 (0.239)	0(1)	-0.271(5.165)	> <b>100 (0.081)</b>	0.021(0.552)
70	0.319 (0.579)	0.664 (0.293)	<b>4.16 (0.04)</b>	-0.325(4.094)	> <b>100 (0.002)</b>	-0.008(0.509)
122	0.320 (0.446)	0.529 (0.212)	0 (1)	-0.231(6.131)	> <b>100 (0.044)</b>	0.019 (0.546)
127	0.076 (0.628)	-0.059(0.960)	-1.00 (0.32)	-0.430(1.955)	> <b>100 (0.031)</b>	-0.086(0.359)
137	0.464 (0.331)	<b>1.127 (0.037)</b>	<b>1.60 (0.07)</b>	-0.045(20.786)	> <b>100 (0.084)</b>	0.147 (0.820)
166*	<b>0.808 (0.094)</b>	<b>1.464 (0.079)</b>	0 (1)	-0.200 (7.530)	> <b>100 (0.007)</b>	0.148 (0.776)
223	0.640 (0.199)	<b>1.345(0.094)</b>	<b>3.14 (0.09)</b>	-0.203 (7.385)	> <b>100 (0.100)</b>	0.124 (0.749)
238	0.711 (0.254)	<b>1.361 (0.095)</b>	0.67 (0.63)	0.222 (6.739)	>100 (0.141)	0.142 (0.764)
253	0.410 (0.392)	1.114 (0.344)	1.59 (0.31)	-0.316 (3.990)	> <b>100 (0.078)</b>	0.020 (0.550)
257	0.746 (0.158)	<b>1.255 (0.094)</b>	1.66 (0.13)	-0.191 (7.933)	> <b>100 (0.081)</b>	0.129 (0.749)
306	0.283 (0.640)	0.574 (0.455)	0 (1)	-0.393 (3.000)	> <b>100 (0.044)</b>	-0.051 (0.433)
320	0.300 (0.800)	0.610 (0.110)	<b>3.10 (0.01)</b>	-0.100 (12.150)	>100 (0.170)	0.240 (0.586)
321	0.606 (0.248)	<b>1.212 (0.099)</b>	1.43 (0.15)	-0.188 (8.085)	>100 (0.158)	0.106 (0.726)
ShxB						
6	0.355 (0.477)	0.694 (0.239)	1.00 (0.22)	-0.204 (0.104)	> <b>100 (0.082)</b>	0.035 (0.608)
10	0.650 (0.161)	<b>1.330 (0.059)</b>	<b>2.71 (0.03)</b>	-0.111 (3.173)	> <b>100 (0.079)</b>	0.171 (0.812)
30*	0.423 (0.343)	<b>0.865 (0.054)</b>	<b>0.94 (0.10)</b>	-0.137 (0.318)	> <b>100 (0.027)</b>	0.137 (0.827)
31*	0.269 (0.534)	0.442 (0.726)	0.79 (0.74)	-0.223 (3.634)	<b>33.127 (0.052)</b>	0.052 (0.635)
41*	0.500 (0.360)	0.690 (0.190)	<b>1.44 (0.10)</b>	-0.170 (20.490)	>100 (0.210)	0.060 (0.387)



47*	0.227 (0.446)	0.291 (0.297)	0 (1)	-0.329 (0.0010)	>100 (0.018)	-0.001(0.490)
105	0.603 (0.127)	<b>1.148 (0.082)</b>	<b>3.55 (0.10)</b>	-0.132( 1.401)	>100 (0.086)	0.141 (0.777)
110	0.091 (0.641)	-3.203 (0.593)	-2.01 (0.79)	<b>1.082 (347.479)</b>	>100 (0.100)	-1.979 (0.524)
121	0.368 (0.292)	0.489 (0.248)	2.34 (0.03)	-0.195 (0.012)	>100 (0.094)	0.037 (0.598)
138	0.598 (0.185)	<b>1.287 (0.077)</b>	<b>2.46 (0.04)</b>	-0.129 (29.164)	>100 (0.100)	0.160 (0.782)
163	-0.449 (0.915)	-1.351 (0.292)	0.66 (0.79)	-0.754 (0.192)	>100 (0.002)	-0.269 (0.199)
169	0.471 (0.188)	1.018 (0.072)	0 (1)	-0.136 (0.832)	>100 (0.094)	0.109 (0.755)
171	0.245 (0.509)	0.897 (0.270)	1.73 (0.17)	-0.227 (0.248)	>100 (0.011)	0.020 (0.592)
184	<b>0.828 (0.098)</b>	<b>2.739 (0.027)</b>	<b>2.90 (0.10)</b>	<b>1.350 (498.23)</b>	>100 (0.036)	<b>0.635 (0.950)</b>
189	0.208 (0.650)	0.416 (0.382)	0.91 (0.26)	-0.301 (0.021)	>100 (0.100)	-0.013 (0.500)
235	0.189 (0.648)	0.352 (0.478)	0 (1)	-0.612 (0.705)	>100 (0.044)	-0.056 (0.427)
266	0.500 (0.250)	0.630 (0.140)	<b>1.88 (0.04)</b>	-0.359 (0.005)	>100 (0.210)	0.060 (0.363)
288	0.380 (0.360)	0.950 (0.450)	<b>4.34 (0.08)</b>	-0.120 (0.130)	>100 (0.410)	0.010 (0.330)
295	0.586 (0.224)	<b>1.315 (0.089)</b>	<b>1.98 (0.08)</b>	-0.119 (3.548)	>100 (0.112)	0.164 (0.809)
330	0.528 (0.251)	<b>0.913 (0.039)</b>	0 (1)	-0.136 (0.324 )	>100 (0.056)	0.140 (0.841)
335	0.237 (0.675)	1.781 (0.243)	1.87 (0.36)	<b>0.491 (89.472)</b>	>100 (0.266)	0.216 (0.826)
347	0.360 (0.460)	0.910 (0.210)	<b>2.78 (0.05)</b>	-0.110 (0.390)	>100 (0.230)	0.720 (0.499)
348	0.341 (0.296)	0.454 (0.193)	0.63 (0.20)	-0.193 (0.005)	>100 (0.027)	0.040 (0.610)
<b>ShxC</b>						
23	0.345 (0.762)	0.367 (0.500)	1.62 (0.23)	-0.332 (2.313)	>100 (0.037)	-0.034 (0.455)
41	2.219 (0.129)	2.863 (0.056)	2.02 (0.55)	-0.075 (2.616)	>100 (0.075)	<b>0.562 (0.952)</b>
44	1.260 (0.232)	<b>1.196 (0.051)</b>	1.23 (0.12)	0.259 (10.205)	>100 (0.070)	0.147 (0.865)
47	0.042 (0.733)	0.662 (0.581)	0 (1)	-0.224 (1.939)	>100 (0.077)	0.015 (0.529)
135	1.631 (0.133)	1.242 (0.093)	1.26 (0.18)	0.090 (4.801)	>100 (0.022)	0.133 (0.822)
137	0.305 (0.580)	1.236 (0.578)	-1.76 (0.36)	-0.517 (0.176)	>100 (0.054)	0.081 (0.637)
157	0.729 (0.562)	0.838 (0.305)	0 (1)	-0.124 (2.420)	>100 (0.079)	0.030 (0.614)
169	1.014 (0.289)	<b>1.294 (0.099)</b>	0 (1)	0.090 (4.856)	>100 (0.122)	0.115 (0.778)
175	0.050 (0.720)	-1.30 (0.37)	<b>2.89 (0.09)</b>	0.110 (1.090)	>100 (0.230)	0.630 (0.762)
186	1.069 (0.277)	1.222 (0.166)	0 (1)	-0.023 (3.060)	>100 (0.005)	0.097 (0.722)
210	-0.395 (0.733)	0.021 (0.993)	8.00 (0.18)	-0.870 (0.000)	>100 (0.043)	0.024 (0.493)
235	0.940 (0.360)	10.08 (0.21)	<b>3.25 (0.07)</b>	0.260 (0.360)	>100 (0.670)	0.275 (0.773)
236	0.538 (0.533)	1.594 (0.546)	2.83 (0.35)	-0.402 (0.634)	>100 (0.021)	-0.026 (0.563)

239	1.846 (0.228)	2.039 (0.113)	4.44 (1.2)	-0.092 (2.479)	>100 (0.012)	0.286 (0.869)
241	0.838 (0.476)	0.985 (0.238)	6.20 (0.08)	-0.076 (2.672)	>100 (0.013)	0.068 (0.694)
<b>ShxD</b>						
11	0.500 (0.210)	1.75 (0.16)	<b>3.43 (0.09)</b>	0.00 (0.060)	>100 (0.180)	0.120 (0.574)
35	0.183 (0.625)	0.807 (0.422)	0 (1)	-0.286 (2.979)	>100 (0.089)	-0.033 (0.521)
49	0.219 (0.458)	0.967 (0.383)	-0.62 (0.43)	0.038 (18.633)	>100 (0.010)	0.048 (0.616)
54	0.324 (0.301)	<b>1.249 (0.085)</b>	0 (1)	0.064 (14.321)	>100 (0.100)	0.108 (0.776)
71	0.406 (0.360)	1.467 (0.444)	0 (1)	-0.184 (8.112)	>100 (0.091)	-0.043 (0.621)
73	0.224 (0.524)	1.835 (0.559)	0 (1)	-0.277 (5.105)	>100 (0.075)	0.001 (0.557)
105	0.422 (0.679)	2.286 (0.400)	4.70 (0.25)	-0.188 (9.068)	>100 (0.025)	0.068 (0.712)
109	0.216 (0.444)	0.704 (0.274)	1.81 (0.15)	-0.163 (3.079)	>100 (0.048)	0.019 (0.558)
167*	0.348 (0.386)	1.200 (0.545)	2.22 (0.43)	-0.165 (9.386)	<b>36.519 (0.047)</b>	0.131 (0.688)
205	0.193 (0.617)	0.689 (0.377)	1.72 (0.24)	-0.260 (0.390)	>100 (0.089)	0.001 (0.539)
215	0.290 (0.470)	1.41 (0.22)	<b>4.79 (0.07)</b>	-0.258 (2.597)	>100 (0.110)	0.124 (0.348)
<b>Zen</b>						
40	<b>3.531(0.059)</b>	<b>3.231(0.023)</b>	7.60(0.01)	<b>0.678 (421.199)</b>	>100(0.033)	0.597(0.942)
45	3.781(0.300)	4.249(0.191)	3.12(0.29)	<b>0.357 (126.438)</b>	>100(0.032)	0.536(0.869)
117	1.808 (0.241)	<b>1.823 (0.099)</b>	0 (1)	-0.104 (17.316)	>100(0.123)	0.162 (0.736)
123	0.636 (0.556)	0.296 (0.842)	1.27 (0.61)	-0.401 6.696	<b>35.467 (0.068)</b>	-0.060 (0.417)
160	2.021 (0.321)	1.032 (0.626)	7.47 (0.13)	<b>0.145 (69.002)</b>	6.294 (0.192)	0.123 (0.644)
165	1.496 (0.309)	<b>1.327 (0.070)</b>	4.45 (0.01)	-0.123(7.274)	<b>35.467 (0.083)</b>	0.137 (0.761)
167	1.741 (0.256)	1.815 (0.100)	0.98 (0.10)	-0.113 (16.665)	>100 (0.039)	0.162 (0.733)
170	1.114 (0.434)	0.333 (0.830)	-1.07 (0.64)	-0.394 (8.135)	>100 (0.039)	-0.024 (0.488)
218	<b>4.546 (0.026)</b>	<b>5.130 (0.007)</b>	0 (1)	<b>0.745 (669.09)</b>	>100 (0.000)	<b>1.330 (0.994)</b>
226	<b>5.896 (0.010)</b>	<b>2.237 (0.095)</b>	3.04 (0.48)	-0.140 (15.559)	>100 (0.051)	0.181 (0.699)
366	1.997 (0.269)	1.630 (0.357)	1.61 (0.52)	0.324 (110.623)	2.572 (0.347)	0.288 (0.760)
538	0.540 (0.667)	0.484 (0.384)	2.07 (0.14)	-0.475 (0.392)	>100 (0.032)	-0.045 (0.409)

## Appendix III

### Appendix III Note1 – Primers and cycling conditions for nuclear genes\*.

Gene	PCR Primers (5' > 3')	Sequencing Primer (5' > 3')
<i>wingless</i>	GAGTGCAAATGCCACGGTATGTCTGG ACTTCGCAGCACCAGTGGAAACGTGCA	GAGTGCAAATGCCACGGTATGTCTGG
<i>zerknüllt</i> (exon 1)	CAGCAAGTACCGATCAATAC AAGAAGTGATTTGGGCAATG	CAGCAAGTACCGATCAATAC
<i>zerknüllt</i> (exon 2 and 3)	GGTTGTTCTGCGGTAAATTAC ACTATCAGAGACCCGGTATG CCTGCAATAGA AACTCGGTTAC	ACTATCAGAGACCCGGTATG GTATTTGCTCGTTGTAGTGG
<i>ShxA</i>	CCTGCAATAGA AACTCGGTTAC TTAGCTAAGCCTCGTGAATATG	CCTGCAATAGA AACTCGGTTAC TTAGCTAAGCCTCGTGAATATG
<i>ShxB</i> (exon1)	AACCGCAAAGTCTCATGTCTAC ACAATATCACCCGGCAGTTAC	AACCGCAAAGTCTCATGTCTAC
<i>ShxB</i> (exon2)	TCACTCTTAATCCAGATGTAC GGAGGATTACTAAATGGTATTG	GGAGGATTACTAAATGGTATTG
<i>ShxC</i>	TTGGAAGCTTTGCCACGTTA ACACAAACGATCGTGACTCT	TTGGAAGCTTTGCCACGTTA
<i>ShxD</i>	CGCCGGCACTTCAATCCTT CCCTTGCATCCCATTGCTTTTAC	CCCTTGCATCCCATTGCTTTTAC

\*The *wg* and *zen* genes were amplified from total genomic DNA extracted from dried butterflies using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions. Double-stranded DNA was amplified through polymerase chain reactions (PCR) in 20- $\mu$ L volumes containing: 12.4  $\mu$ L DEPC treated H<sub>2</sub>O, 4 $\mu$ L 5x iProof HF buffer, 0.4  $\mu$ L 10 mM dNTPs, 0.5  $\mu$ L of each primer (10  $\mu$ M), 0.2  $\mu$ L iProof DNA Polymerase (Bio-Rad, 2U/  $\mu$ L) and 1  $\mu$ L of extracted DNA.

*Wingless* PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 65°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Forward primer was used for sequencing.

*Zerknullt* was amplified using two separate primer pairs, one covering the first exon and one covering the second and third exon. PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 61°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Forward and reverse primers were used for sequencing of exons 2 and 3, and forward primer for exon1.

*ShxA* was amplified using one set of primers. PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 64°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Sequencing was performed using both forward and reverse primers.

*ShxB* was amplified using two separate primer pairs, one covering the first exon and one covering the second. PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 60°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Forward primer was used for sequencing of exon1, and reverse primer for exon2.

*ShxC* was amplified using one set of primers. PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 61°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Sequencing was performed using forward primer.

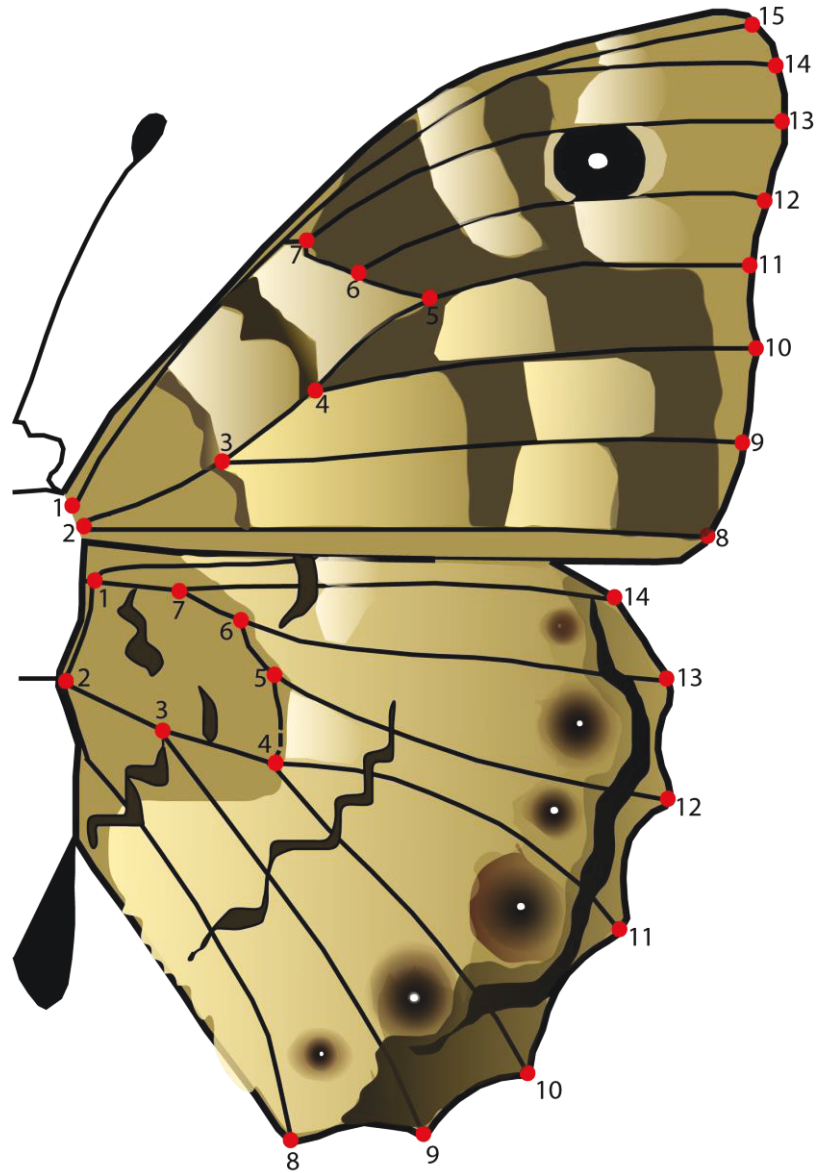
*ShxD* was amplified using one set of primers. PCR conditions consisted of an initial cycle at 98°C for 60s, followed by 40 cycles of 98°C for 10s, 61°C for 30s, 72°C for 90s, and a final extension at 72°C for 5min. Sequencing was performed using reverse primer.

PCR products showing a clear band in gel electrophoresis were purified using QIAquick PCR Purification columns (Qiagen) and sequenced by EurofinsGenomics Ltd. Sequences were edited and aligned using the MUSCLE algorithm implemented in Ugene (Okonechnikov et al., 2012).

**Appendix III Table 1 – *Wolbachia* screen**

<b>Female</b>	<b>Location</b>	<b>Island</b>	<b><i>Wolbachia</i> Presence</b>
27	Bavella	Corsica	<b>Yes</b>
24	Bavella	Corsica	N/A
25	Cavallo Morto	Corsica	<b>Yes</b>
15	Zonza	Corsica	N/A
16	Tempio Pausania	Sardinia	<b>Yes</b>
18	Zonza	Corsica	<b>Yes</b>
19	Bavella	Corsica	<b>Yes</b>
21	Zonza	Corsica	N/A
22	Zonza	Corsica	<b>Yes</b>
1	Aritzo	Sardinia	N/A
2	Aritzo	Sardinia	<b>No</b>
3	Aritzo	Sardinia	N/A
4	Aritzo	Sardinia	N/A
5	Aritzo	Sardinia	N/A
6	Aritzo	Sardinia	<b>No</b>
7	Aritzo	Sardinia	<b>Yes</b>
8	Sualeddu	La Maddalena	<b>Yes</b>
9	Aritzo	Sardinia	<b>No</b>
10	Tempio Pausania	Sardinia	<b>Yes</b>
11	Tempio Pausania	Sardinia	N/A
12	Aritzo	Sardinia	N/A
13	Asco	Corsica	<b>Yes</b>
17	Bavella	Corsica	N/A
20	Bonifacio	Corsica	<b>No</b>
23	Desulo	Sardinia	<b>No</b>
26	Solenzara	Corsica	<b>Yes</b>
28	Bavella	Corsica	<b>Yes</b>
29	Bavella	Corsica	N/A
30	Asco	Corsica	<b>Yes</b>
31	Aritzo	Sardinia	<b>Yes</b>
32	Pietralba	Corsica	<b>Yes</b>

## Appendix IV



**Appendix IV - Figure 1. Ventral side of the *Pararge aegeria* wing surface highlighting landmarks used for the geometric morphometrics.**

Homologous landmarks used in the geometric morphometric analysis are shown on both forewing and hindwings. Internal landmarks used were numbers 3-6, all other landmarks correspond to external positions. Figure adapted from Breuker *et al.*,(2007).

*Pararge aegeria* - WT



*Pararge aegeria* - *WntA* CRISPR



**Appendix IV - Figure 2. *Pararge aegeria* wild-type and *WntA* mKO mutants.** Dorsal and ventral Wild type *P. aegeria* (top) and of injected individuals (bottom) showing clear *WntA* phenotypes are shown. Rate of mosaicism leads to a range of effects in pigmentation defects.

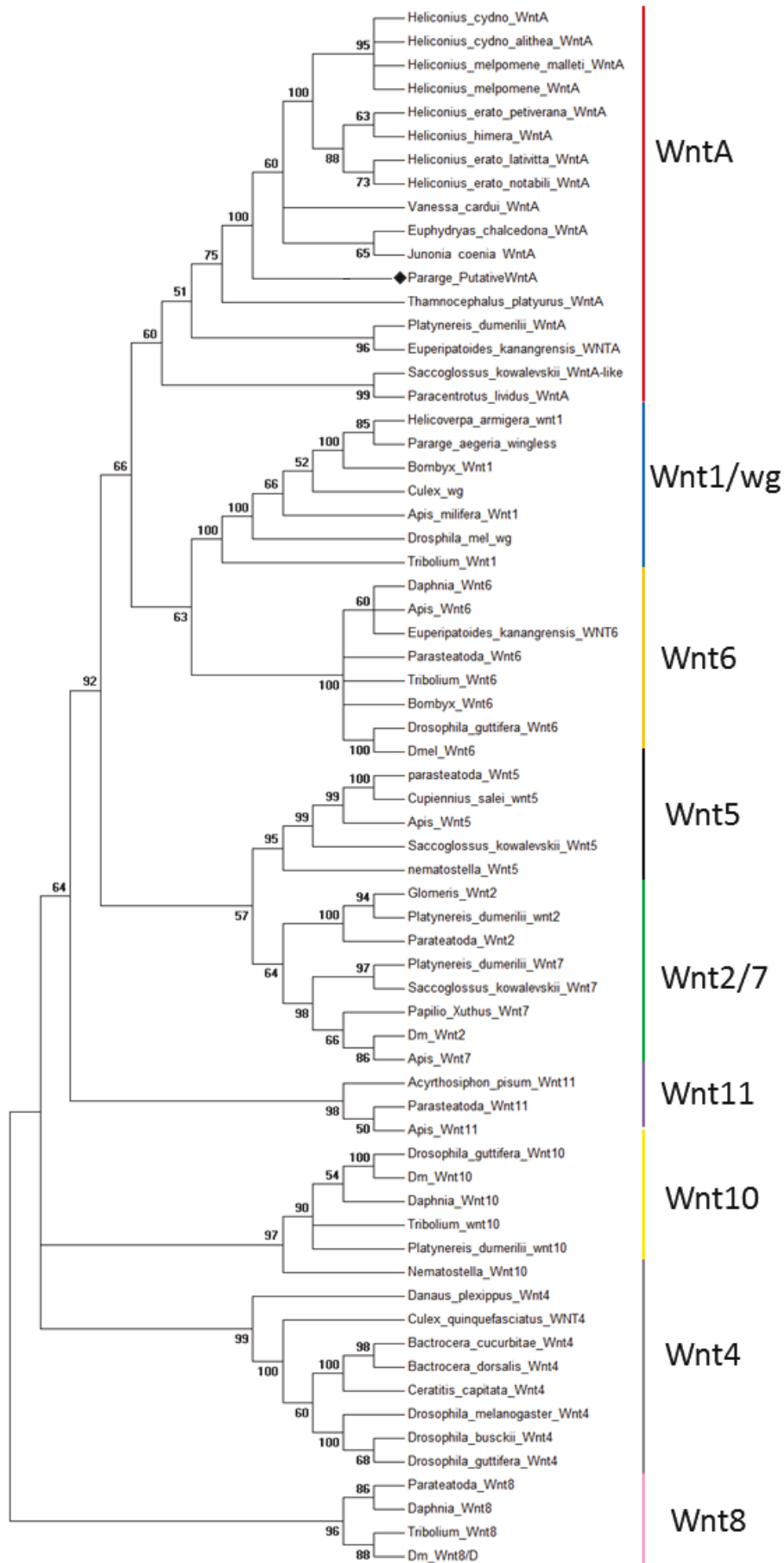
*Pararge aegeria* - WT



*Pararge aegeria* - Yellow CRISPR



**Appendix IV - Figure 3. *Pararge aegeria* wild-type and yellow mKO mutants.** Dorsal and ventral Wild-type *P. aegeria* (top) and of injected individuals (bottom) showing clear *yellow* phenotypes are shown. Rate of mosaicism leads to a range of effects in pigmentation defects.













**Appendix IV – Figure 4. Phylogenetic analysis of *WntA* in *P. aegeria*.**

The evolutionary history was inferred by using the Maximum Likelihood method based on the Le and Gascuel (2008) model. The model was chosen based on the automatic model prediction tool implemented in MEGA7 (Kumar et al., 2016). The tree with the highest log likelihood is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (+G, parameter = 0.9355)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 4.0795% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 76 amino acid sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. All branches with less than 50% bootstrap support are collapsed. There were a total of 138 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). The position of the *P. aegeria WntA* gene is highlighted with a rhombus in the tree.

**Effect of Treatment on Shape**

	Internal		External	
	FW 	HW 	FW 	HW 
♀	ns	ns	ns	ns
♂	**	*	**	**

**Assymetry of Shape (left-right)**

	Internal		External	
	FW 	HW 	FW 	HW 
♀	***	ns	ns	ns
♂	***	ns	***	***

**Appendix IV – Figure 5. Summary of significant effects reported on wing shape.**

Geometric morphometric results for wing shape showing significant effects for overall shape (top) and shape asymmetry (bottom) split between males and females for internal and external landmarks. See main text and appendix IV Table 2 for details. ns = not significant; \* significant at  $p < 0.05$ ; \*\* significant at  $p < 0.01$ ; \*\*\* significant at  $p < 0.001$ .

## Appendix IV – Additional file 1. CRISPR/Cas9 Protocol for *Pararge aegeria*

### Reagents

Qiagen PCR purification kit (Qiagen)  
RNeasy RNA purification kit – Qiagen  
MAXIscript T7 Kit (Life technology)  
IProof high fidelity polymerase (Bio-Rad)

### Design of sgRNA target sequence

SgRNA target sites were designed by seeking sequences corresponding to N20NGG on exon regions of the sense or antisense strand of the target gene using the E-CRISP program. Candidate target sequences are then BLASTed against the *P. aegeria* genome to eliminate those with potential off-target sites using strict criteria, where the candidate editable site is defined only when the seed region (12 nucleotides (nt) to protospacer adjacent motif (PAM) NGG) is unique. From candidate editable sites, we selected those with the first two bases of GG, for sgRNA synthesis using a T7 promoter.

### sgRNA Preparation

1) Order a unique oligonucleotide encoding T7 polymerase binding site and the sgRNA target sequence N20 (CRISPRF:

GAAATTAATACGACTCACTATAN20GTTTTAGAGCTAGAAATAGC) and common oligonucleotide encoding the remaining sgRNA sequence (sgRNAR: AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCT TATTTAACTTGCTATTTCTAGCTCTAAAAC).

2) Primer self-amplification of CRISPRF and sgRNAR

ddH<sub>2</sub>O 67 µl  
5X iProof Buffer 20 µl  
10 mM dNTPs 2 µl  
10 µM CRISPRF 5 µl  
10 µM sgRNAR 5 µl  
iProof DNA polymerase 1 µl  
100 µl

98 °C 30 s  
98 °C 10 s  
60 °C 30 s × 35 cycles  
72 °C 15 s  
72 °C 10 min

Purify with Qiagen PCR purification kit, and elute in 30 µl EB.  
Measure concentration using NanoDrop.  
Expected concentration should be around 150ng/ul.

### In vitro transcription of sgRNA using Megashortscript T7 kit

(1) Thaw the frozen reagents, and then keep 4 nucleotides and enzyme on ice but keep 10X Reaction Buffer at room temperature.  
(2) Assembly transcription reaction at room temperature and incubate

Nuclease-free water \*  
DNA template \*\*

10X Reaction Buffer 2  $\mu$ l  
10 mM ATP 1  $\mu$ l  
10 mM CTP 1  $\mu$ l  
10 mM GTP 1  $\mu$ l  
10 mM UTP 1  $\mu$ l  
Enzyme Mix 2  $\mu$ l  
20  $\mu$ l

(\*==Calculate based on DNA template concentration; \*\*== 1  $\mu$ g linearized PMD18-T7-sgRNA plasmid or 600 ng purified self-amplification PCR-product)

Mix thoroughly and incubate at 37°C for 4-6 h.

- (3) Add 1  $\mu$ l TURBO DNase, mix well, and incubate at 37°C for 15 min.
- (4) Add RNase-free water to the DNase I-treated transcription reaction up to 50  $\mu$ l.
- (5) Add 5  $\mu$ l 5 M Ammonium Acetate and vortex to mix.
- (6) Add 3 volumes 100% ethanol.
- (7) Chill the solution at -20°C for 30 min or longer.
- (8) Spin for >15 min at maximum speed in a 4°C centrifuge.
- (9) Carefully discard the supernatant, and wash the pellet once with cold 70% ethanol.
- (10) Elute in 10  $\mu$ l RNase-free water, and determine sgRNA(s) concentration by NanoDrop. We recommend a concentration of more than 1  $\mu$ g/ $\mu$ l for later use.
- (11) Store sgRNA (s) in 5  $\mu$ l aliquots at -70°C.

### **Preparation of CRISPR injection mixes**

- (1) Resuspend Cas9 protein at 1 $\mu$ g/ $\mu$ L in 0.15% Phenol Red/ RNase-free Water. If you ordered 50 $\mu$ g : add 35 $\mu$ L H<sub>2</sub>O and 15 $\mu$ L Phenol Red Stock Solution (0.5%)
- (2) Mix well by vortexing 30s, spin down.
- (3) Prepare 1.5 $\mu$ L aliquots and store for long-term at -80C (recommended: use 0.5mL Eppendorf)
- (4) My injection mixes have a total volume of 7.5 $\mu$ L (Cas9 at 333ng/ $\mu$ L); add sgRNA to a final concentration of 150ng/ $\mu$ L (around 1 $\mu$ L) and RNase-free water for 7.5 $\mu$ L total
- (5) Dispatch in 2-2.5 $\mu$ L aliquots and store injection mixes at -80C until day of injection

### **Eggs**

- (1) add host plant to cage and start timer.
- (2) collect eggs by hand at t<2hrs and transfer to Eppendorf (do not seal!). The goal is to inject at a syncytial stage that will generate a lot of mosaics to obtain “escapers” (non-embryonic lethal effects). Trying different stages is recommended, and keep in mind that staging will depend on temperature. For embryonic phenotypes, injecting <1hr seems reasonable.
- (3) Prepare an injection slide: Align eggs with anterior pole facing outwards on a slide covered with double sided sticky tape using a hair tool or paintbrush.

### **Injection procedure**

- (1) Keep your injection mix aliquot on ice during the injection procedure.
- (2) Load a small amount of CRISPR injection mix into your needle. Use an Eppendorf Femtotip mounted on a 20 $\mu$ L pipette.
- (3) Set the Needle on the Micromanipulator at about 45 degrees. Set your injector to about 30-45psi, injection time 10-20ms.
- (4) “Balance” the positive pressure to evacuate the system pressure.
- (5) Check if your needle easily goes into eggs. You need to puncture the egg quite harshly on the lateral periphery, around the micropyle. These injections can seem quite brutal, it is common to hear the needle poking the eggs with a popping sound. I like to have the needle at a fixed position and move the plate up and down with my left hand (better control of the force)

- (6) Inject outside of an egg and see how the fluid droplet behaves: if it goes back into the needle, you must slightly increase the positive pressure (balance knob) ; if it keeps inflating, you should slightly reduce the positive pressure. A good droplet is no bigger than 5% of the egg volume, and stays at the tip of the needle without inflating.
- (7) Start injecting. After you poke the egg, you can move the needle back to make more room inside, Inject once when the needle seems unobstructed, or several times to push yolk out from the needle until you are sure you have transferred some amount of fluid At first there does not seem to be a perfect way to do this step, so I recommend to experiment with the procedure.
- (8) You will often see a droplet of egg fluid coming out, which immediately seems to polymerize and coat the egg
- (9) Shifting from egg to egg, you will often see yolk in your needle. You could try to push it out in water before moving to the next egg, or just move to the next egg and inject it out, it does not seem to be a problem.
- (10) Eventually, your needle will break and inject larger volumes, In which case you must either decrease injection time or just change needle. I routinely inject 50-100 eggs with one needle.

### **Hatching**

- (1) Remove injected eggs from slide and drop them onto a petri dish containing filter paper using a paintbrush. Be careful not to damage embryos.
- (2) Place petri dish containing injected embryos into incubator at 23C and 65% relative humidity.
- (3) Wait until hatching and observe phenotypes.

**Appendix IV - Table 1. sgRNA target sequences and genotyping primers.**

<b>Gene</b>	<b>sgRNA Name</b>	<b>Target sequence (5' &gt; 3')</b>	<b>Genotyping primers (5' &gt; 3')</b>
<i>WntA</i>	Pa_WntA_sgRNA1	GGCGTCGTTTACCTCAGCTG	CCACGGCACTGGCAGTGGGG
	Pa_WntA_sgRNA2	GGGTGTCGATTGCTTTGCTG	CATCTGCATTTTTCTTCGTG
<i>yellow</i>	Pa_yellow_sgRNA1	GGTCCGGCATGAAATAGCTG	ACACTGATCGGCGTATTAGAAGA
	Pa_yellow_sgRNA2	GGTATAAGTGCTTCTAATAT	ACCACGGGTTATTGTTGGTGA
<i>ShxA</i>	Pa_ShxA_sgRNA1	GGTTTTGCTTCCATTCCAA	TAGCGCCAGTGGCATTTACA
	Pa_ShxA_sgRNA2	GGAAGCAAAACCTCAAGTGG	TCGGGGTAGTCTTCCCATGT
<i>ShxC</i>	Pa_ShxC_sgRNA1	GGAACGTCGCAAAGAGCTCG	TAATTGCTGGAAAACCGAC
	Pa_ShxC_sgRNA2	GGCTCAATGTAAGTGTGGTA	GAAACCAAACCTTGACGCAT

**Appendix IV - Table 2. Asymmetry of overall wing size and shape for internal and external landmarks in forewings and hindwings.**

Centroid size is the dependent variable according to model described by Palmer and Strobeck (1986). Denominators used to calculate *F*-values are the Mean squares of each term below the first. Asymmetry of shape is based on Procrustes sum of squares as a measure of overall variation in shape (see methods). A permutation test was used to determine the statistical significance of each effect. \* significant at  $p < 0.05$ ; \*\* significant at  $p < 0.01$ ; \*\*\* significant at  $p < 0.001$ .

<b>Female Forewing External</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	112.291	56.145	2	10.07***
	Individual	144.944	5.575	26	67.41***
	Side	0.00145	0.00145	1	0.02
	Ind*Side	2.316	0.0827	28	1.91*
	Measurement	2.469	0.0433	57	
<i>Shape</i>					
	Treatment	0.00420	0.000131	32	1.4
	Individual	0.0391	0.0000153	416	4.02***
	Side	0.000661	0.0000413	16	1.77
	Ind*Side	0.0105	0.0000234	448	4.89***
	Measurement	0.00436	0.0000478	912	

<b>Female Forewing Internal</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	2.712	1.356	2	4.28*
	Individual	8.236	0.317	26	37.35***
	Side	0.0124	0.0124	1	1.46
	Ind*Side	0.237	0.00848	28	3.21***
	Measurement	0.151	0.00264	57	
<i>Shape</i>					
	Treatment	0.0104	0.000867	12	0.82
	Individual	0.165	0.00106	156	2.49***
	Side	0.0112	0.00186	6	4.38***
	Ind*Side	0.0713	0.000424	168	4.01***
	Measurement	0.0362	0.000106	342	

<b>Male Forewing External</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	216.819	108.409	2	22.99***
	Individual	273.500	4.716	58	30.88***
	Side	0.0915	0.0915	1	0.6
	Ind*Side	9.164	0.153	60	4.84***
	Measurement	3.847	0.0315	122	
<i>Shape</i>					
	Treatment	0.0132	0.000412	32	2.13**
	Individual	0.180	0.000194	928	4.48***
	Side	0.00157	0.0000978	16	2.26***
	Ind*Side	0.0415	0.0000432	960	8.67***
	Measurement	0.00973	0.0000498	1952	

<b>Male Forewing Internal</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	14.055	7.027	2	25.35***
	Individual	16.077	0.277	58	24.15***
	Side	0.0000790	0.0000790	1	0.01
	Ind*Side	0.686	0.0114	60	5.11***
	Measurement	0.274	0.00224	122	
<i>Shape</i>					
	Treatment	0.0294	0.00245	12	2.53**
	Individual	0.337	0.000969	348	2.83***
	Side	0.0240	0.00399	6	11.67***
	Ind*Side	0.123	0.000342	360	2.33***
	Measurement	0.1088	0.000147	732	

<b>Female Hindwing External</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	77.603	38.802	2	9.32***
	Individual	124.939	4.165	30	71.44***
	Side	0.154	0.154	1	2.65
	Ind*Side	1.865	0.0583	32	2.63***
	Measurement	1.462	0.0222	66	
<i>Shape</i>					
	Treatment	0.00887	0.000317	28	1.53
	Individual	0.0870	0.000207	420	2.77***
	Side	0.00468	0.000334	14	4.47
	Ind*Side	0.0335	0.0000747	448	4.55***
	Measurement	0.0152	0.0000164	924	

<b>Female Hindwing Internal</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	1.299	0.649	2	4.63*
	Individual	4.205	0.140	30	19.81***
	Side	0.0695	0.0695	1	9.83**
	Ind*Side	0.226	0.00707	32	4.09***
	Measurement	0.114	0.00173	66	
<i>Shape</i>					
	Treatment	0.0244	0.00203	12	1.3
	Individual	0.280	0.00156	180	2.93***
	Side	0.00279	0.000465	6	0.88
	Ind*Side	0.102	0.000530	192	2.6***
	Measurement	0.0808	0.000204	396	

<b>Male Hindwing External</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	167.528	83.764	2	21.16***
	Individual	253.382	3.959	64	57.61***
	Side	0.230	0.230	1	3.35
	Ind*Side	4.536	0.0687	66	4.5***
	Measurement	2.045	0.0153	134	
<i>Shape</i>					
	Treatment	0.0122	0.000436	28	1.65**
	Individual	0.237	0.000265	896	3.35***
	Side	0.00403	0.000288	14	3.65***
	Ind*Side	0.0729	0.0000788	924	5.88***
	Measurement	0.0251	0.0000134	1876	

<b>Male Hindwing Internal</b>					
<i>Overall wing size</i>	Effect	SS	MS	df	<i>F</i>
	Treatment	10.137	5.06928	2	25.25***
	Individual	12.850	0.200775	64	24.95***
	Side	0.0264	0.026431	1	3.28
	Ind*Side	0.531	0.008048	66	3.67***
	Measurement	0.294	0.002193	134	
<i>Shape</i>					
	Treatment	0.0694	0.00578	12	2.95*
	Individual	0.753	0.00196	384	3.32***
	Side	0.00249	0.000415	6	0.7
	Ind*Side	0.234	0.000590	396	3.08***
	Measurement	0.154	0.000191	804	